9-6-2018 1:45 PM

Statistical Modeling of CO2 Flux Data

Fang He
The University of Western Ontario

Supervisor
Murdoch, Duncan
The University of Western Ontario Co-Supervisor
Kulperger, Reg
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Carbon dioxide (CO_2) flux is important for agriculture and carbon cycle studies. Only a small proportion of the land is currently covered by proper equipment to directly collect CO_2 flux data. The CO_2 flux data has an obvious annual cycle with the phase changing from year to year. How to build a model to estimate the annual effect and seasonal dynamics is a challenging task. With the help of the Moderate Resolution Imaging Spectroradiometer (MODIS) which is carried by NASA satellites, corresponding data, such as normalized difference vegetation index (NDVI), is freely available from NASA. Our goals are modeling the seasonal dynamics to generate reasonable predictions, and building a model using MODIS data to predict the CO_2 flux data at any location.

In modeling single sites, we treat each year as a multivariate observation. We use functional data analysis (FDA) to smooth the CO_2 flux data for each year separately. Then we use the landmark registration to standardize the seasonal dynamics. On the registered time scale, we build a model of the curves using a multivariate normal distribution. We use Dirichlet regression to model the seasonal dynamics and map the registered curves back to the natural time scale. These steps allow us to simulate CO_2 flux on the natural time scale.

For our spatial study, we study three models. In the first model, we decompose the CO_2 flux data into different components, and build a model based on the spatial correlations of each component. The second model is a functional linear regression model (FLRM), where we use NDVI as the covariate. Both CO_2 flux and NDVI are annual multivariate observations treated as functional objects. In the third model, we use a generalized additive model (GAM) to analyze the data as a time series indexed by day, with covariates such as NDVI, latitude, longitude etc.

We use parametric bootstrap to validate our single location modeling on 55 flux sites. The local and the majority of global coverage rate are around 95%. Among the three spatial models, the GAM performed best in that it had the lowest out of sample prediction mean square error. The FLRM also shows a great potential for modeling with limited information.

Keywords: CO_2 flux, NDVI, phase variation, landmark registration, spatial modeling, FDA, GAM

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ecologists study the interactions of organisms and their relationships with their physical environment. One branch of ecology studies the interactions between the living and nonliving environments, including the movement of carbon between the soil, organisms, and the atmosphere (Krasny, 2003). This is the scientific motivation for this thesis. Carbon dioxide (CO_2) movement is a large part of the studies, as it is involved in respiration (emitting CO_2) and photosynthesis (absorbing it). The flow of CO_2 can be locally measured; this is called the CO_2 flux. Amongst other objectives, ecologists are interested in prediction of CO_2 flux, and hence in its modeling. One of the challenges comes from the phase variation of the cycle of the CO_2 flux data due to seasonal dynamics. This problem will become clear and intuitive after we present an example of the CO_2 flux data in Section 1.2.3.

Often, two sources of data are used in ecosystem studies, the first of which is remote sensing data measured from a distance by aircraft, radar or satellite, while the other is ground-based data obtained from fixed stations located in a specific area. An example of remote sensing data is MODIS (Moderate Resolution Imaging Spectroradiometer) data. MODIS is an instrument currently on board two satellites. The ground-based data used in our study are carbon dioxide (CO_2) flux data, which measures how much carbon dioxide moves through a unit area per unit time. CO_2 flux is an important component of the total atmospheric carbon balance. It is impor-

tant to the study of global climate change. Positive records of CO_2 flux occur during periods when the plants release more CO_2 than the amount they absorb, and records are negative when more CO_2 is absorbed by photosynthesis than is released by respiration.

Some MODIS products are collected only every eight days or even less frequently, which limits our ability to study the phase variation (Kuenzer et al., 2015). The frequency of collecting CO_2 flux data is every 30 minutes or hourly. For spatial coverage, MODIS data is worldwide, while the land coverage of CO_2 flux data is low, and it is costly to get ground-based data at a new location.

In this thesis, we build models for CO_2 flux data. We start with single location modeling. The model uses only CO_2 flux data, and helps to study the phase variation. For multi-location modeling, we propose three models. In the first model, we decompose the CO_2 flux data to different components. The model relies on the spatial correlation of each component. The second model uses functional linear regression. We use one MODIS product as a covariate. Different from the ordinary linear regression, both the dependent and independent variables are functional objects, whose advantage will be explained later in Section 2.2.3. The third model is a generalized additive model, which has the flexibility to combine multiple covariates.

In this thesis we chose four sites in Saskatchewan to conduct our single location modeling. For the multi-location study, we used a subset of the sites in North America, and more than 150 sites were involved. For every site, we used CO_2 flux data from ground stations and corresponding NDVI (Normalized Difference Vegetation Index) data from MODIS. We detail the data to a greater extent in Sections 1.1 and 1.2. Section 1.1 provides the background information on the flux and remote sensing data, including the description of the measuring instruments and the specific data products. Section 1.2 presents the data acquisition and filtering process used in this study.

1.1 Background

In this section, we explore ground-based and remote sensing data in greater detail.

Section 1.1.1 presents general information on ground-based flux towers and their geographical distribution.

Section 1.1.2 briefly describes NASA's Earth Observing System (EOS) and its two special missions, Terra and Aqua, both of which carry MODIS. In the end, we describe one MODIS product NDVI.

1.1.1 Ground-based Data

CO_2 flux can be directly measured using a flux tower or chamber. Three examples of flux towers are shown in Figure 1.1. This equipment works locally and can be installed at a fixed location on the earth. However, in the area between flux towers, the flux information is unknown.

(a) Flux tower in forests. (de Araújo, 2015)

(b) Flux towers at MSU's Kellogg Biological Station. (Krasean, 2013)

(c) 6400-09 Soil CO_2 flux chamber. (LI-COR, 2016)

Figure 1.1: Flux tower examples. Image (a) shows a flux tower in forests. It is able to measure below/above-canopy CO_2 flux. Image (b) presents one type of flux tower used in fields. Image (c) displays a flux chamber used to measure soil CO_2 flux.

In 2014, there were over 650 flux tower sites in the world. They are plotted in Figure 1.2. Worldwide, the distribution of the flux tower networks is not uniform. A large proportion of

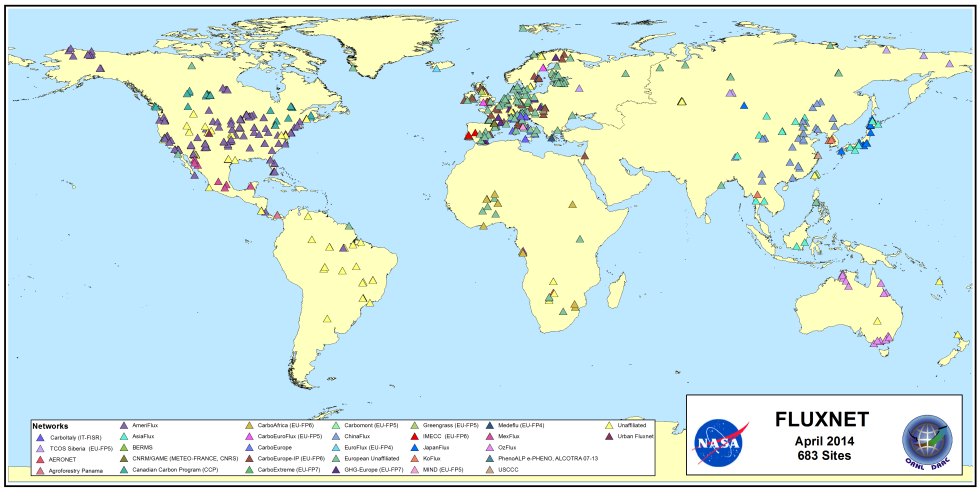


Figure 1.2: Distribution of flux tower networks. (FLUXNET, 2014)

the land remains uncovered.

1.1.2 Remote Sensing Data

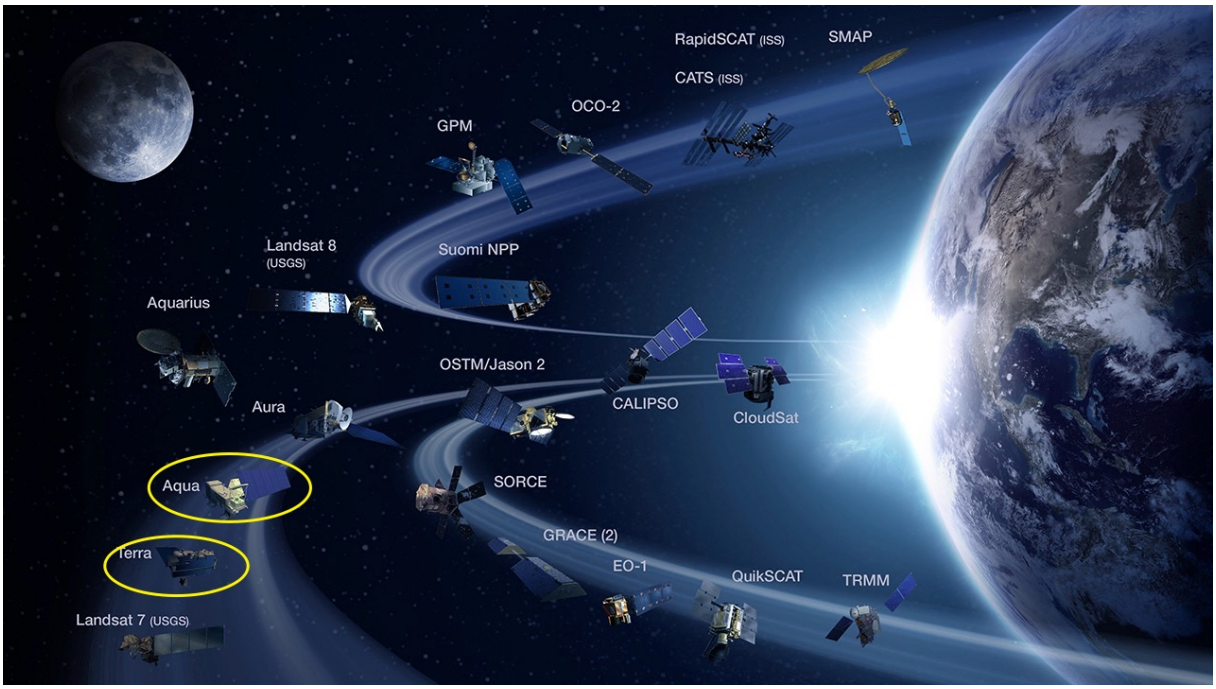


Figure 1.3: NASA earth science division operating missions. (NASA, 2016)

MODIS is an instrument on board the Terra and Aqua satellites, which are two EOS mis-

sions launched in 1999 and 2002 respectively. According to NASA (2016) "EOS is a coordinated series of polar-orbiting and low inclination satellites for long-term global observations of the land surface, biosphere, solid earth, atmosphere, and oceans". Figure 1.3 shows Terra, Aqua and other NASA earth science division operating missions. Table 1.1 summarizes information on Terra and Aqua compiled by Parkinson et al. (2006).

Table 1.1: Some details of Terra and Aqua satellites.

	Primary Mission	**Launched**	**Passes equator**	
			Direction	**Time**
Terra	atmosphere, land, and oceans	Dec 18, 1999	North to South	Morning
Aqua	water, radiative energy flux, aerosols, etc.	May 4, 2002	South to North	Afternoon

MODIS has 36 spectral bands, with wavelengths ranging from 0.4 μm to 14.4 μm. The resolutions are 1 *km* or finer (2 bands at 250 *m*, 5 bands at 500 *m*, 29 bands at 1 *km*). MODIS's operational altitude is 705 *km*, from which vantage point it scans the whole world every one to two days.

To better view how Terra and Aqua work, Figure 1.4 displays the orbit of Terra, and Figure 1.5 plots the swath of Aqua. Both satellites are in sun-synchronous orbits, which means that each of the satellites passes over any given point of the planet's surface at the same local solar time. The gap between adjacent swaths will be covered by the next round of scan.

NASA releases the MODIS data, such as NDVI, leaf area index, and thermal anomalies and fire, free of charge to the public at `https://modis.ornl.gov/data.html`.

The raw MODIS data cannot be used directly. It needs decoding, interpretation, scaling, and positioning. NASA processes the raw MODIS data in different forms and stages known as products. Table 1.2 shows definitions of MODIS products from level 0 to level 4.

NDVI

Because one of our goals is to predict ground-based data by means of remote sensing data, the two types of data we choose should be related. The ground-based data we chose is related

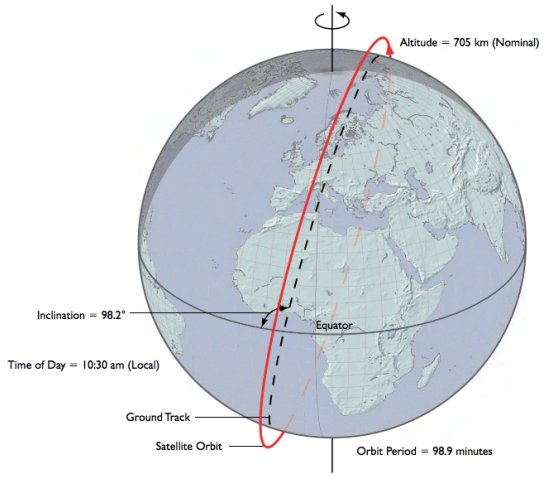


Figure 1.4: 3D model of Terra Satellite orbit.(King et al., 2007)

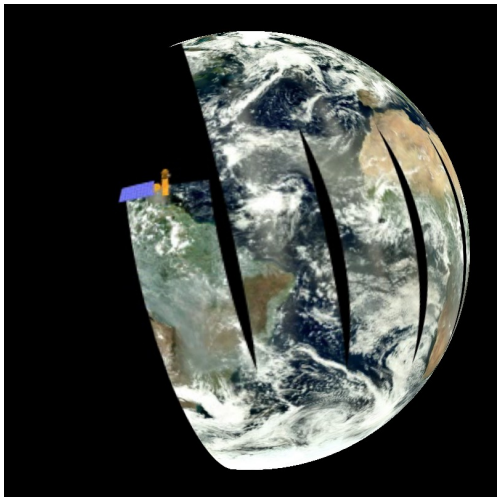


Figure 1.5: Aqua satellite and MODIS swath. (NOAA, 2016)

Table 1.2: Definitions of MODIS data processing levels.

Product Levels	**Product Definitions**
Level 0	Unprocessed instrument data.
Level 1A	Unprocessed instrument data alongside ancillary information.
Level 1B	Data processed to sensor units, e.g. brightness temperatures.
Level 2	Derived geophysical variables, e.g. sea ice concentration.
Level 3	Variables that are mapped on a grid, e.g. data using EASE-Grid.
Level 4	Modeled output or variables derived from multiple measurements.

to the greenness of plants. Therefore, among various MODIS products, we chose the NDVI for the study. The greenness of plants is affected by the growth stage and the health condition. Since CO_2 flux from plants is also related to these two things, NDVI can serve as a predictor of flux. How NDVI is calculated better explains why it is related to the greenness of plants. Generally speaking, greener vegetation reflects more near-infrared (NIR) light and less visible (VIS) light. On the contrary, less green or sparse vegetation reflects more VIS light and less NIR light. A vegetation index is an indicator of greenness, and a common one is the NDVI, which is defined as:

$$\text{NDVI} = \frac{\text{NIR} - \text{VIS}}{\text{NIR} + \text{VIS}}, \tag{1.1}$$

where NIR and VIS are the proportions of respective light reflections, NIR and VIS are between $[0, 1]$, and NDVI is within $[-1, 1]$. NDVI is defined by a ratio of two quantities with the same measurement unit and therefore NDVI itself has no units.

Table 1.3: NDVI to object(s) reference table.

NDVI	**Possible Object(s)**
below 0.1	barren areas of rock, sand or snow
around [+0.2, +0.3]	shrub and grassland
(+0.6, +0.8)	temperate and tropical rainforests

When NDVI is given, we can roughly infer the corresponding object based on the values in Table 1.3 (Weier and Herring, 2018).

Although NDVI is widely used in quantitative assessments, it is highly sensitive to factors including the thickness of clouds, soil-moisture content, etc. (Hogrefe et al., 2017).

1.2 Data

In this section, we focus on describing the data used in the study. Section 1.2.1 introduces the data sources and the acquisition process. Section 1.2.2 describes how we filtered our data

due to the inconsistent data collection frequency. Section 1.2.3 displays the filtered data for a representative site.

1.2.1 Data Sources

Flux towers give researchers the ability to monitor carbon dioxide flux as it varies between the earth and atmosphere, and signal increases or decreases in the gas. The flux data for our single location study (Chapters 3 and 4) are from the Canadian Carbon Program (CCP) at `https://daac.ornl.gov/daacdata/fluxnet/FLUXNET_Canada/data/`, while the flux data for our multi-location study (Chapter 5) are from AmeriFlux at `http://ameriflux.lbl.gov/`. The remote sensing data are from NASA. Table 1.4 summarizes different data sources and their relationship with the information introduced in Section 1.1.

Table 1.4: Data sources summary.

Category	Type	Product	Data Source
ground based	flux tower data	CO_2 flux	CCP and AmeriFlux
remote sensing	MODIS data	NDVI	ORNL DAAC

In the rest of Section 1.2.1, we elaborate on each data source.

1. **CCP**

   CCP operates 32 sites wherein various types of ecosystems are monitored with different variables and timeframes and provides more than ten different types of observations from each site. What we are interested in is the CO_2 flux above-canopy data. We wanted our single location study CO_2 flux data to meet the following requirements:

   (a) Each site has at least three years of high-quality data with an annual pattern.

   (b) Sites are geographically close to each other.

   (c) Each site has only one species, but we want multiple species over the selected sites.

We checked all of the CCP sites. Table A.1 in Appendix A lists18 sites which recorded CO_2 flux above-canopy data. The time frames listed in Table A.1 are the time ranges that each site was under maintenance, however, the actual time ranges of collecting CO_2 flux data are sometimes much shorter than their overall maintenance time, and vary greatly from site to site.

When all of the rules were applied, we ended up with four sites in Saskatchewan as listed in Table 1.5. Their locations are shown in Figure 1.6. We plot the sites in both satellite and terrain maps because the satellite map pictures the landscapes better and terrain map gives us the ability to discern the specific locations.

Table 1.5: CO_2 flux above-canopy data information.

Longitude	Latitude	Site Name	Time Range	Data Type
-106.1978	53.6289	SK Old Aspen	2003 - 2008	CO_2 Flux Above Canopy 39m
-104.6920	53.9163	SK Old Jack Pine	2003 - 2008	CO_2 Flux Above Canopy 28m
-105.1178	53.9872	Sk Southern Old Black Spruce	2003 - 2008	CO_2 Flux Above Canopy 25m
-104.6453	53.8758	Sk 1975 (Young) Jack Pine	2004 - 2006	CO_2 Flux Above Canopy 16m

2. **AmeriFlux**

Besides CCP, another source of high-quality flux data is AmeriFlux, which contains over 150 sites across the Americas measuring gases, heat, wind, soil, and so on. Figure 1.7 displays the locations of all AmeriFlux sites.

The variable we are interested in is "FC", which stands for CO_2 turbulent flux (μmol CO_2 $m^{-2}s^{-1}$). Unlike the CO_2 flux of the four selected sites collected from CCP which have only above/below canopy CO_2 flux, the data collected from AmeriFlux include above/below canopy CO_2 flux, and soil CO_2 flux. The number of sites we sampled was 158. After checking, site "CR-Lse" did not provide any "FC" data, so we eliminated it from our study. For each site, we collected information on latitude, longitude, elevation, and vegetation type. For more information about the AmeriFlux data variables please see `http://ameriflux.lbl.gov/data/aboutdata/data-variables/`.

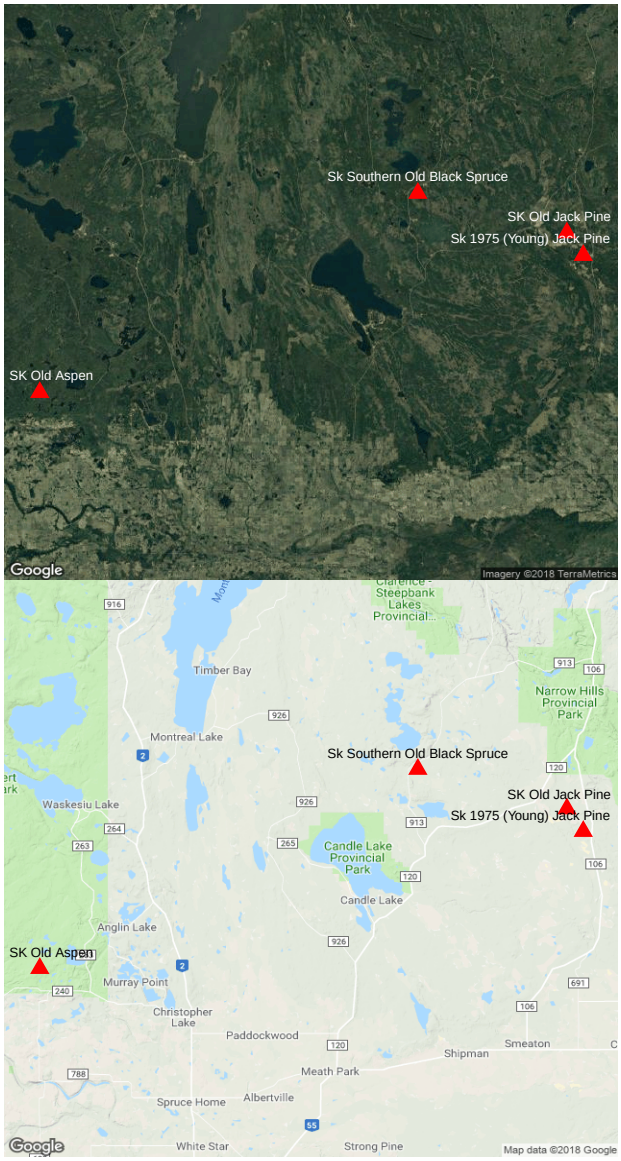


Figure 1.6: Sites of single location study.

There are eleven different vegetation types across the sites we sampled. Their abbreviations and full names are listed in Table 1.6.

From the bar chart in Figure 1.8, we can tell that the number of sites is unevenly distributed across species.

3. **ORNL DAAC**

MODIS data used in this thesis was collected from the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) MODIS land product subsets. The

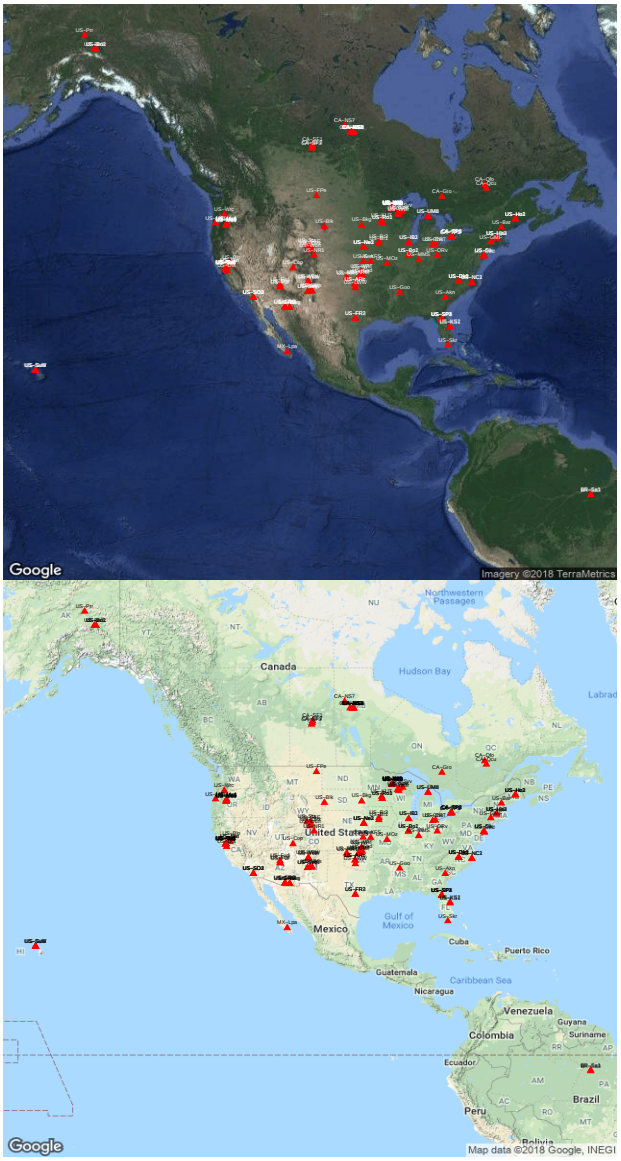


Figure 1.7: Locations of sampled AmeriFlux sites.

product we chose was named "MOD13Q1 Vegetation Indices (NDVI, EVI)", which is a Level 3 product with 250 *m* resolution. The requested data area is approximately 250 *m* wide and 250 *m* long.

1.2.2 Data Filtering

Since the NDVI data are highly sensitive to weather and other conditions, the 16-day average NDVI data are not exactly the average over 16 days, but NASA may choose the most

Table 1.6: Vegetation types' abbreviations and full names.

Abbreviation	Full Name
CRO	Croplands
CSH	Closed shrublands
DBF	Deciduous broadleaf forests
EBF	Evergreen broadleaf forest
ENF	Evergreen needleleaf forest
GRA	Grasslands
MF	Mixed forest
OSH	Open shrublands
SAV	Savannas
WET	Permanent wetlands
WSA	Woody savannas

desirable record during that time to represent it. The frequencies of collecting CO_2 flux are either every 30 minutes or hourly. It is always easier to analyze two different types of data with the same frequency, so we rendered the flux data uniform with the satellite data using the following process:

1. We calculated the daily average of CO_2 flux data.

2. We determined the 16-day average of the daily average of CO_2 flux data. Since 365 is not a multiple of 16, $365 = 16 \times 22 + 13$, the first 22 16-day average data periods were each calculated by the 16 daily average data, while the last period was calculated using only the 13 daily average data.

1.2.3 Data View

We plot the CO_2 flux data of SK Old Aspen in Figure 1.9. The data presents an obvious annual pattern, which is a slightly increasing trend from January to April, scarcely decreasing

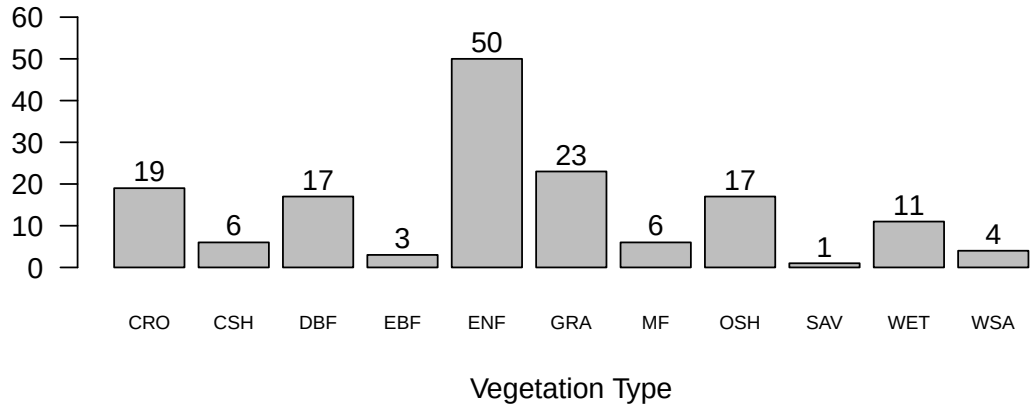


Figure 1.8: Bar chart of the number of AmeriFlux sites in each vegetation type.

trend from October to December, and a sharp decreasing trend followed by an acute increasing trend from April to October each year. The NDVI data for SK Old Aspen is shown in Figure 1.10. Its pattern shows a peak in the summer and trough in the winter.

Since we have so many sites for the multi-location study, we do not present them one by one. In Chapter 5 we will plot some of them to illustrate the data filtering.

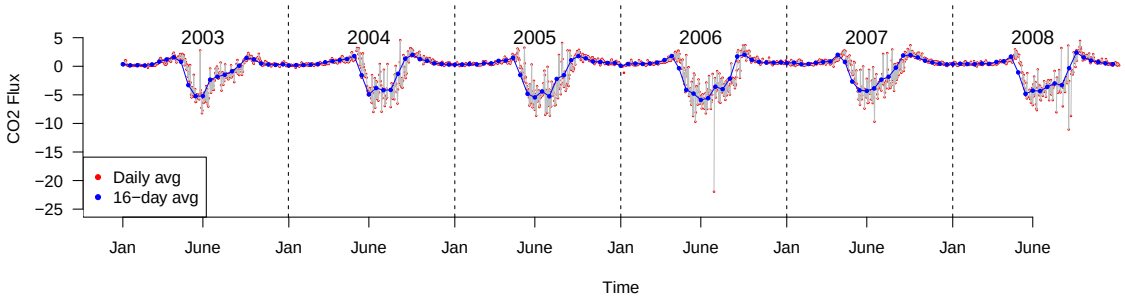


Figure 1.9: Saskatchewan old aspen, latitude 53.63°N and longitude 106.20°W, CO_2 flux data from 2003 to 2008. The red dots are the daily average data and the blue dots are the 16-day average data.

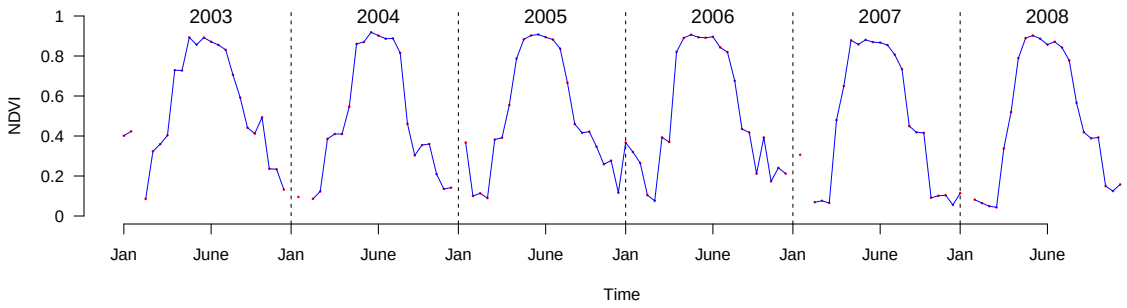


Figure 1.10: Saskatchewan old aspen NDVI data from 2003 to 2008.

1.3 Outline of the Thesis

From Figure 1.9, we can see there is a pattern repeated year after year, but the cycles are not identical. In each year, they start and stop at slightly different times, i.e. there are variations in the phases. Figure 1.10 shows that NDVI data have a negative correlation with the CO_2 flux data. Biologically this makes sense because, in the summer, trees are greener, and thus they absorb more CO_2.

Compared to the satellites Terra and Aqua, which have worldwide coverage and have been continuously working for at least 15 years, the flux towers work only locally, their maintenance may not be continuous and the length of available time is comparatively short. All of these qualifications add challenges to the related CO_2 flux studies. Our objectives in this thesis are to address these problems in the following ways:

1. Model the phase variation of CO_2 flux.
2. Generate reasonable CO_2 flux predictions for observed sites.
3. Predict CO_2 flux data at any location by means of NDVI and other covariates, such as location, vegetation type and elevation.

In Chapter 2, we briefly review some of the literature in ecology and statistics. Two preliminary statistical analyses, using a generalized additive model (GAM) and functional data analysis (FDA), are presented together with a discussion of their limitations for solving our

problems. For this reason, a new model is required.

We present our single location modeling in Chapters 3 and 4. This model addresses objectives one and two above. In Chapter 3, we describe the structure and the components of our single location model, followed by a description of its random structure. Chapter 4 is an expansion of Chapter 3, and focuses on the theoretical details of the four steps: registration under certain constraints, simulation in the registered time scale using the multivariate normal (MVN) distribution, mapping the simulated curves back to the real-time scale, and gauging the measurement error. The chapter ends with the model checking method and its corresponding results.

Chapter 5 introduces our spatial model. This model helps with the third objective. We consider three models, in each case starting with the details of data preparation, followed by details of modeling. Then we show the results. The discussion of each model ends with a short summary. At the end of this chapter, we include an overall discussion.

We draw our conclusions on each model in Chapter 6 and present our future work.

Three additional tables are listed in Appendix A. They are tables of the 18 CCP sites, 38 MODIS products and the selected 89 AmeriFlux sites.

The software used in our study is R (R Core Team, 2017).

Chapter 2

Literature Review

The MODIS data have been collected since 2000, and around 40 products (Table A.2) are now available. The flux tower data collection has an even longer history. Not only large amounts of data have been recorded, but a number of studies have been carried out using these data and similar data from other sources. We introduce some studies from both ecological and statistical points of view.

Section 2.1 summarizes how ecologists have analyzed the MODIS and flux tower data. Those analyses can be categorized as biomass comparisons, gap filling discussions, etc. Besides regression and other general methods, some studies have applied machine learning techniques, such as a support vector machine.

Section 2.2 focuses on analyzing the data from a statistical perspective. We first review a recent study using a statistical model to estimate the ground CO_2 flux by atmospheric carbon data globally. The model will be introduced step by step together with its connections with our study. Then we illustrate two other statistical methods by performing preliminary analyses at one location.

2.1 Ecology

Ecology studies involving MODIS and flux data cover mainly the following three topics:

1. Comparisons of the biomass calculated using MODIS and flux data. Those studies consider many biomass quantities, such as regional evaporation and gross primary production (GPP), which is the amount of chemical energy produced in a given length of time. Heinsch et al. (2006) and Kalfas et al. (2011) conducted studies to calculate GPP from MODIS and flux data respectively and used linear regression to reveal their relationship. Similar analyses have been applied to evaporation (Cleugh et al., 2007).
2. The quality of MODIS and flux tower data is sensitive to weather conditions or unexpected mechanical defects, and therefore gap filling techniques are required. Moffat et al. (2007) compared gap filling techniques for carbon flux data.
3. Phenology studies have used both MODIS and flux data. Klosterman et al. (2014) discussed using three different sigmoid functions to estimate phenology dates and compared the data-driven outcomes with visual assessment. This study also summarized the challenges inherent to phenology, including setting up standard protocols to choose suitable biological characteristics and comparing current methods of transition date estimation, such as the start of spring and the end of fall.

In addition to the above three topics, machine learning techniques, for example, a continental scale model built by a support vector machine, have been used in the prediction of GPP (Yang et al., 2007) and evapotranspiration (Yang et al., 2006).

2.2 Statistics

From a statistical perspective, a number of methods can be applied to study data with an annual pattern. In Section 2.2.1, we review a hierarchical model from a recent study. This model sequentially processes the global atmospheric carbon data to estimate the ground CO_2 flux. In Section 2.2.2, we treat the data as a single time series and apply GAM. In Section 2.2.3, we list the general FDA analysis steps with a brief example. Next, we focus on the details of spline smoothing and a registration method.

2.2.1 A Hierarchical Model - A Review of A Recent Study

Cressie (2018) presents a hierarchical statistical model using remote sensing data alone to study the geographic distribution of sources and sinks of CO_2. In a hierarchical model, the outcomes from one level will be used in the next level.

The first level in Cressie (2018) is to calibrate raw CO_2 data, which are processed from three spectrometers on board the satellite OCO-2, into an approximately 2200-dimensional vector of radiances denoted as $\boldsymbol{Y}$.

At the second level, for each location, Cressie (2018) uses the radiances to estimate 20 CO_2 values at different altitudes, surface pressure, and aerosols, which are denoted as $\boldsymbol{X}$. The model is

$$\boldsymbol{Y} = F(\boldsymbol{X}) + \epsilon, \tag{2.1}$$

where $\boldsymbol{Y}$ is the calibrated radiances from level one, and F is a known and nonlinear function. This is similar to our single location model (Chapters 3 and 4), where we use ground based CO_2 flux data as $\boldsymbol{Y}$ in (2.1) and will estimate its specific parameters as $\boldsymbol{X}$. Solving $\boldsymbol{X}$ from (2.1) may encounter problems: the solution may not exist, the solution may not be unique, or the solution may not continuously depend on the data. However, a regularized solution can be calculated when considering $\boldsymbol{X}$ associated with a probability distribution. After applying a weighted average to the 20 estimated CO_2 values, a column-averaged CO_2 value is generated. The collection of the averaged CO_2 values across all locations, Cressie (2018) denoted it as $\boldsymbol{Y}_d$, will be used in the next two levels.

Cressie (2018) describes the third level of the model as noise filtering and gap filling, which is to interpolate all the averaged CO_2 data $\boldsymbol{Y}_d$ using kriging. This step is similar to our model of multiple sites (Chapter 5), where we are interested in the CO_2 flux data at unobserved locations.

In the final step, Cressie (2018) models surface fluxes using the processed atmospheric CO_2 data $\boldsymbol{Y}_d$ by conditional expectation. We obtain the estimation of flux data from our multi-location model.

The article also provides methods for conducting inference of each level. We use different methods because the flux tower data allows numeric and graphical validation of our models.

2.2.2 Additive Model

Hastie and Tibshirani (1990) mention "the additive model is a generalization of the usual linear regression model." Let's intuitively explain it using the multiple linear regression model, in which y is the dependent variable, $x_i, i = 1, \ldots, p$ are covariates, and $\beta_j, j = 0, 1, \ldots, p$ are coefficients, and ϵ is the measurement error:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \tag{2.2}$$

where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.

We can generalize (2.2) to the following expression:

$$y = f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \epsilon. \tag{2.3}$$

The same assumptions on ϵ are still made. The special part is that $f_i(x_i), i = 1, \ldots, p$ is now "an arbitrary unspecified function" and "a smooth may be defined as an estimate of $f_i(x_i)$" (Hastie and Tibshirani, 1990). The general options for estimating $f(x_i)$ are kernel smoothing, spline smoothing, etc. A critical property of (2.3) is that the effects of independent variables are additive, which means we can analyze the impact of predictor variables separately.

We use a univariate additive model using R package `mgcv` (Wood, 2001) to model Old-Aspen data retrieved from CCP (Figure 1.9). The covariates are day and NDVI. For the smoothing basis, we chose cyclic cubic regression splines for Day, and used the thin plate regression splines for NDVI. The other model fitting arguments are left as the defaults. The model is

$$\text{Flux} \sim f_1(\text{Day}) + f_2(\text{NDVI}), \tag{2.4}$$

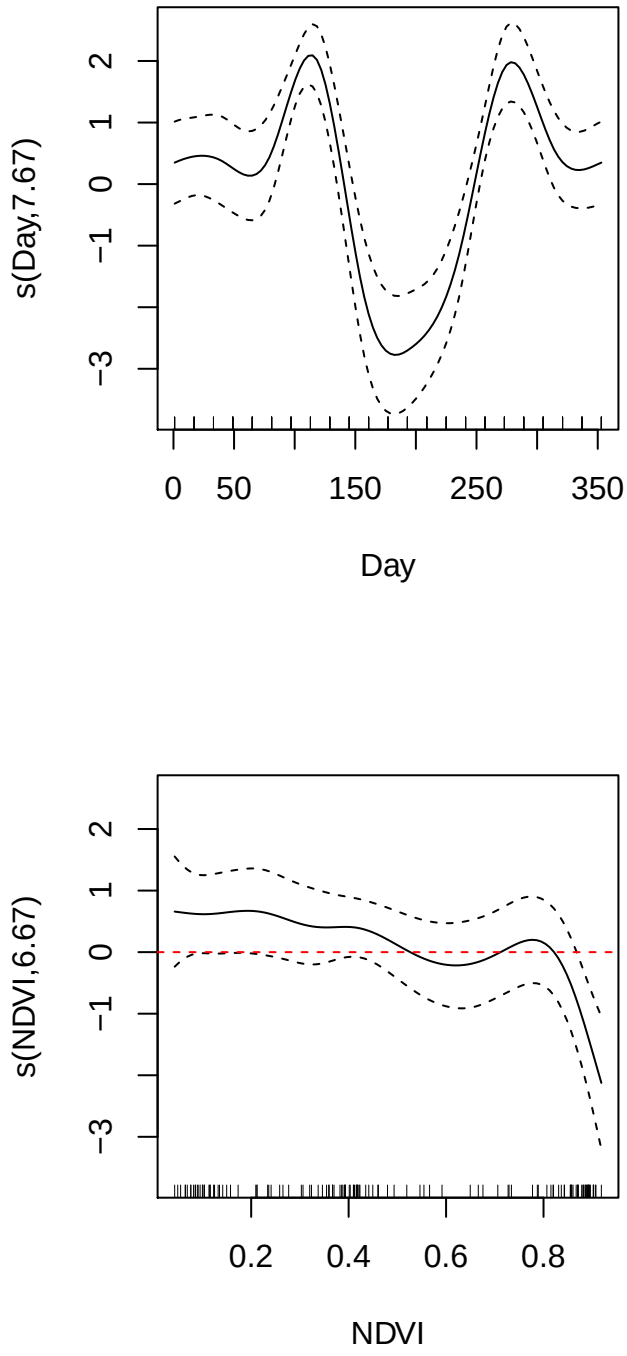


Figure 2.1: GAM estimates of (2.4). The two plots show the smooths in the additive fit. The top plot displays the smooth versus day of year, the bottom plot displays the smooth versus NDVI level. The estimated effects are solid lines with 95% confidence limits shown as dashed lines. Observed values of the covariates are shown as ticks on the bottom axis.

where functions f_1, f_2 are smooth functions; day is a numeric variable starting at the first day of each year increased by every 16 days; NDVI is numeric. The smoothed terms f_1(Day) and f_2(NDVI) are the averages of CO_2 flux by the change of day and NDVI respectively. Throughout this chapter, we use f as a generic function, which will become clear in the context.

We plot the model estimations in Figure 2.1. The estimation of f(Day) gives the annual underlying trend, which shows that independent variable Day has contribution from day 80 to day 300 through out a year. Since the estimated effect of NDVI are mostly horizontal and around 0, we add a reference red dashed line. When NDVI exceeds 0.85, it has an influence on CO_2 flux.

The residuals of the fitting are plotted in Figure 2.2. The residuals are not homogeneous. One possible reason for the inhomogeneity is that each year the phase of the underlying pattern is different. Thus, the assumption that the data at one location has the same cycle over the years is not quite appropriate. Moreover, this model doesn't help with the phase variation study. Methods other than GAM should be applied. One potential tool is FDA which we introduce in the next section.

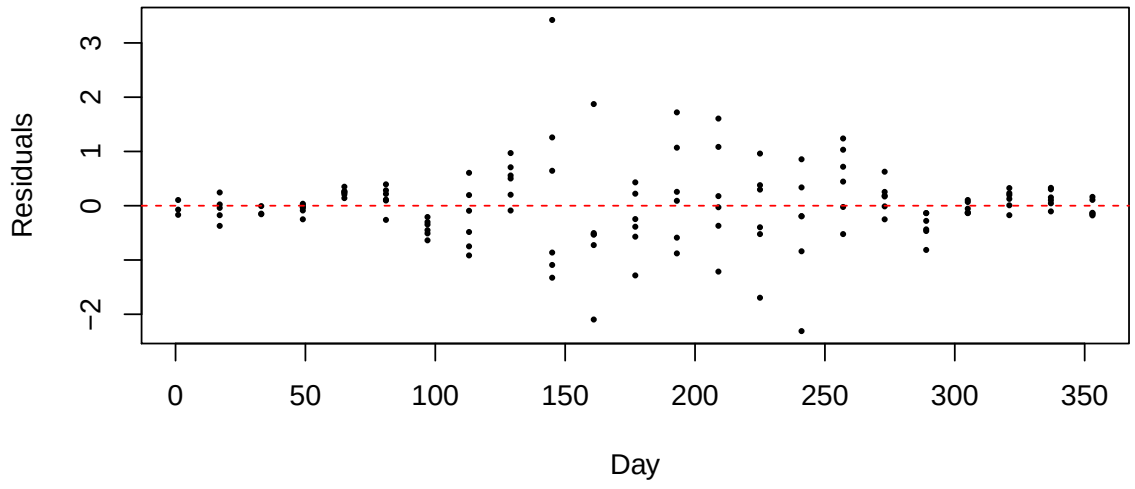


Figure 2.2: Residuals of GAM fit (2.4). The residuals show higher variation in the middle of the year. This violates our assumption of $E(\epsilon) = \sigma^2$.

2.2.3 Functional Data Analysis

Ramsay and Li (1998) say "A functional datum is not a single observation but a set of measurements over a continuum, usually time." One assumption of FDA is that the underlying process generating the data is smooth. We usually start FDA by applying smoothing techniques, then set the original data aside. Further analyses are based on the smoothed curves.

Applications of FDA have appeared in a large number of publications in multiple fields. Epifanio and Ventura-Campos (2011, 2014) published their analysis of applying FDA in bone shape and hippocampal shape analysis by using registration tools and functional principal component analysis (fPCA). Leng and Müller (2006) proposed a method whereby they used functional logistic regression based on functional principal components to classify gene expression curves into known groups. Tian (2010) discussed three statistical challenges of applying functional data analysis in brain imaging studies. Gorrostieta et al. (2014) applied FDA to storm waves.

The steps of using FDA can be roughly summarized as:

1. Use functional smoothing to express each replication as a functional object.
2. Register the smoothed curves.
3. Conduct further analysis. Methods are selected according to the goal of the study. Many studies use functional principal component analysis (fPCA), but our study does not use this.

Figure 2.3 is a functional data smoothing example from the R package `fda` (Ramsay et al., 2018). The heights of 54 girls were taken at unevenly spaced time points from age zero to 18. The black dots in Figure 2.3 (a) represent the original data points. The smooth lines are the fitted curves for each girl. After smoothing, we can estimate a girl's height at any time point between zero and 18. One of the advantages of converting discrete data to a functional object is that we can get the derivatives. The first derivatives, a.k.a. the growth rates are shown in Figure 2.3 (b). The overall trend of growth rate is decreasing; however, from around age nine to 13, the growth rates display an increasing trend. Biologically, that time interval may be consistent

with puberty. The second derivatives (the growth accelerations) are shown in Figure 2.3 (c). There is more fluctuation in each curve, which makes it a little bit harder to identify the overall trend.

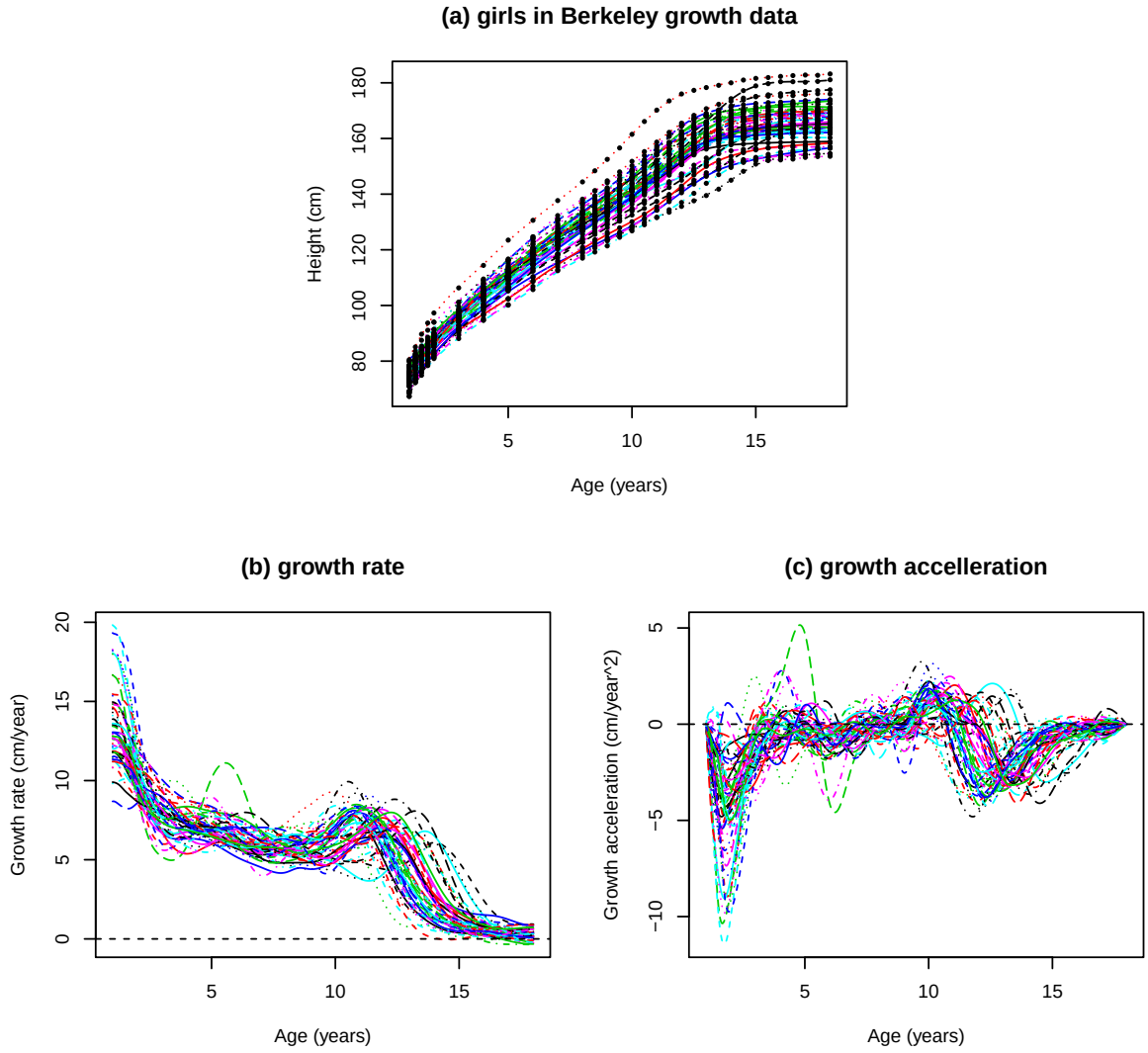


Figure 2.3: Functional data smoothing example of Berkeley girls growth study. In (a), the black dots are the observations and the dashed lines are the smoothed curves. In (b) and (c), we plot the first and second derivative of curves in (a) separately.

Each curve of growth acceleration has a drop from positive to negative between ages 10 and 15. The trend is similar and the difference is the starting time. If we merge this trend together, it is easier to identify the overall trend. In Figure 2.4 (a), we marked the locations where decreasing accelerations hit zero. For demonstration purposes, we only used 34 acceleration

curves. In Figure 2.4 (b), we display the data after pulling all crossings to the same location - their mean value. In this way, the trend is more clear. This feature alignment technique is called data registration.

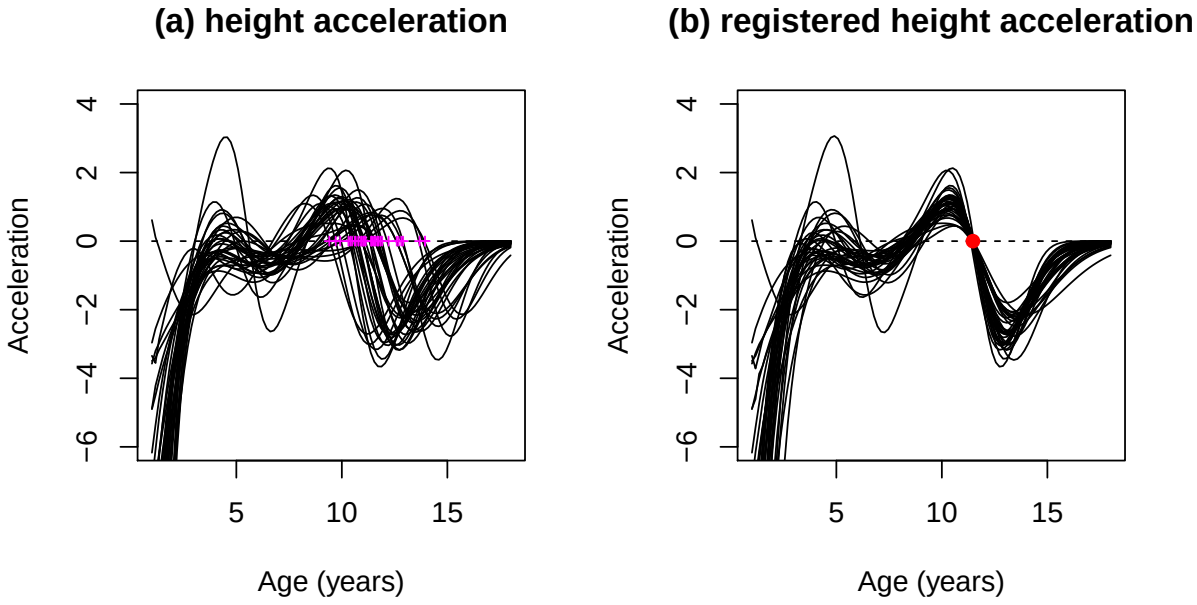


Figure 2.4: Functional data registration example of girls growth data. In (a), we mark the time points of the changes in sign of the smoothed data between age 10 and age 15. In (b), we warp the time scale of each curve, thus, the marked sign change points in (a) are registered at one data point - their average. The registered curves on the right-hand side better reveal the overall underlying trend.

Spline Smoothing

The k^{th} order spline with s interior knots at $t_1 < t_2 < \cdots < t_{s-1} < t_s$ is a function $\phi : \mathbb{R} \mapsto \mathbb{R}$, where ϕ is a continuous piecewise polynomial of degree k on intervals $(-\infty, t_1], [t_1, t_2], \ldots, [t_s, \infty)$, and it also has continuous derivatives at the s knots up to $k-1$ times. The most common case is $k = 3$, the cubic splines, which have continuous first and second derivatives.

The j^{th} observation and its observed day in year i are denoted as y_{ij} and t_{ij} separately, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. Each year for m paired data (t_{ij}, y_{ij}), we fit the regression

model $f(t) = E(Y|T = t)$ using a k^{th} order spline with s interior knots as follows:

$$f(t_{id}) = \sum_{d=1}^{s+k+1} \beta_d \phi_d(t_{id}),$$

where $t_{id} \in [1, 365]$ and $\beta_1, \ldots, \beta_{s+k+1}$ are the coefficients.

The coefficients of year i are estimated using least squares, by minimizing

$$\sum_{j=1}^{m} \left(y_{ij} - f(t_{ij})\right)^2, \tag{2.5}$$

and the fitted line of year i is

$$\hat{f}(t_{id}) = \sum_{d=1}^{s+k+1} \hat{\beta}_{id} \phi_{id}(t_{id}), \tag{2.6}$$

If we write this spline regression in its matrix form, the i^{th} year's coefficient $\boldsymbol{\beta}_i$ can be estimated by the combination of design matrix Φ_i and observation vector $\boldsymbol{y}_i$.

The matrix form of spline regression for the i^{th} year is defined as

$$\boldsymbol{y}_i = \Phi_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \tag{2.7}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})^\intercal$, $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{i,s+k+1})^\intercal$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{i,s+k+1})^\intercal$ and

$$\Phi_i = \begin{pmatrix} \phi_1(t_{i1}) & \phi_2(t_{i1}) & \cdots & \phi_{s+k+1}(t_{i1}) \\ \phi_1(t_{i2}) & \phi_2(t_{i2}) & \cdots & \phi_{s+k+1}(t_{i2}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_{im}) & \phi_2(t_{im}) & \cdots & \phi_{s+k+1}(t_{im}) \end{pmatrix}. \tag{2.8}$$

The estimated value for $\boldsymbol{\beta}_i$ is

$$\hat{\boldsymbol{\beta}}_i = (\Phi_i^\intercal \Phi_i)^{-1} \Phi_i^\intercal \boldsymbol{y}_i. \tag{2.9}$$

The old aspen data after spline smoothing are shown in Figure 2.5, where the vertical dashed lines indicate the locations of manually selected knots, which are summarized in Table

2.1.

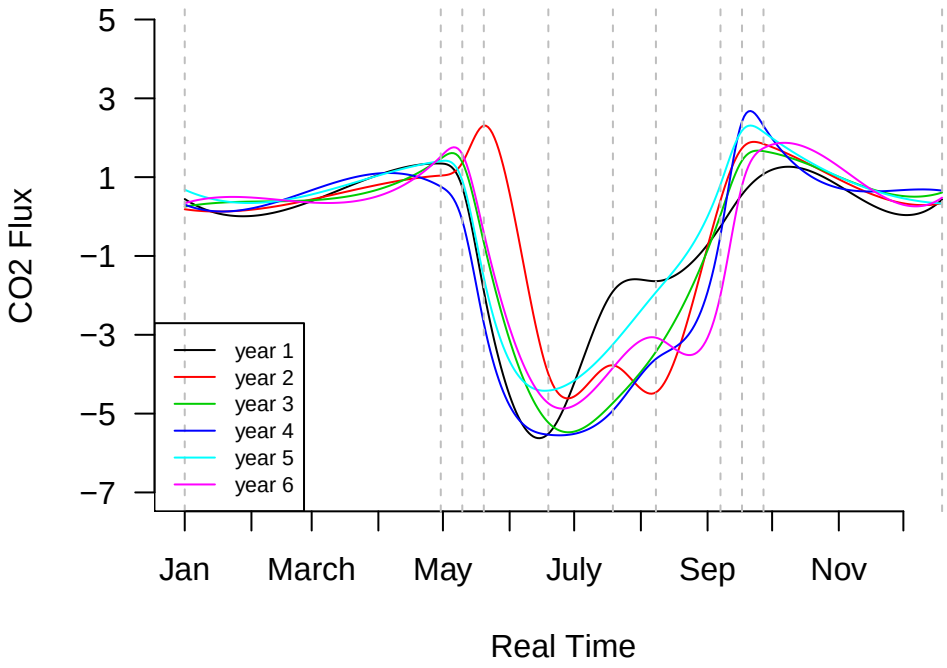


Figure 2.5: Six-year old aspen data after spline smooth only. The vertical dash lines indicate the location of knots. We place relatively more knots at the turning points of our curves to better capture their underlying trends.

Table 2.1: Knots used in old aspen data.

Day of year	1	120	130	140	170	200
Date	Jan 1	Apr 30	May 10	May 20	Jun 19	July 19
Day of year	220	250	260	270	365	
Date	Aug 8	Sep 7	Sep 17	Sep 27	Dec 31	

Registration

Figure 2.5 displays how the occurrence of extreme points in the smoothed data varies each year: the two peaks around May and October, as well as the low point between the two peaks,

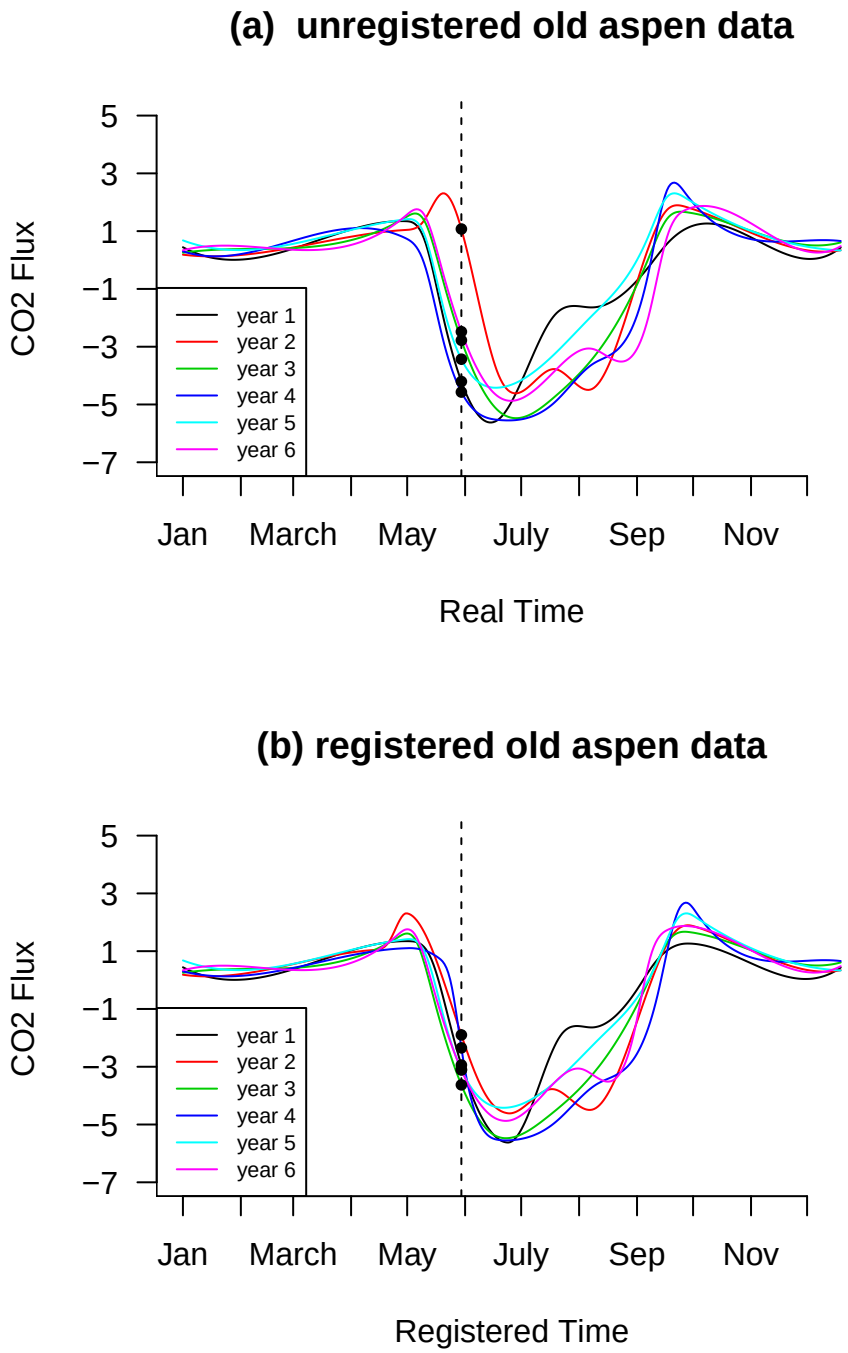


Figure 2.6: Registration effect demonstration. The registered old aspen data are shown on the bottom. We warped the time scale for each year, so that the two peaks and one valley are perfectly aligned. We plot vertical dashed lines at the same time in both images. Points on the registered curves have smaller variation. Because the corresponding warping functions (Figure 2.7) are almost straight lines, we labelled the x-axis of image (b) with months which are proportionally rescaled into the interval $[0, 1]$.

do not line up. This horizontal variation contributes to the observed vertical variation. One tool to reduce the variation is data registration. To demonstrate its effect, Figure 2.6 displays both the unregistered and registered old aspen data. We draw dashed vertical lines at the same time in two cases. At that time point, the solid black dots in Figure 2.6 (b) have much less variance than those in Figure 2.6 (a). The registration method we applied is called landmark registration. Landmarks are typically identified with zeros at the level of some derivatives. With landmarks, we can establish a warping function connecting the real time scale and the registered time scale. Since our study involves the transformation between two different time scales, to distinguish the natural and the registered time scales, we denote them by t and t^* respectively.

Warping Function

Suppose we pick L landmarks $t_{i,l}$ for the i^{th} year. We use $h(t^*)$ to denote the warping function, whose domain is arbitrary, but for convenience, we take it as $t^* \in [0, 1]$. A warping function connects the real time and the registered time:

$$t = h_i(t^*),\ t \in [0, 365] \text{ and } t^* \in [0, 1]. \tag{2.10}$$

In our fitting of the curves in the registered time scale, we use (2.11) to find $f_i^{**}(t^*)$ at $t^* = \frac{1}{365}, \frac{17}{365}, \ldots, \frac{353}{365}$.

$$f_i^{**}(t^*) = f_i(h_i(t^*)),\ \forall i = 1, 2, \ldots, n. \tag{2.11}$$

We use $\boldsymbol{y}_i^*$ to denote the 23 $f_i^{**}(t^*)$ points in the registered time for year i. We fit a smooth $f_i^*(t^*)$ through $\boldsymbol{y}_i^*$ for each year .

Our time warping function has the following properties:

1. $h_i(0) = 0$ and $h_i(1) = 365$.
2. h_i is linear between the consecutive landmarks.
3. h_i is strictly monotonic for $\forall a < b$, $h_i(a) < h_i(b)$. This guarantees the warping function is invertible.

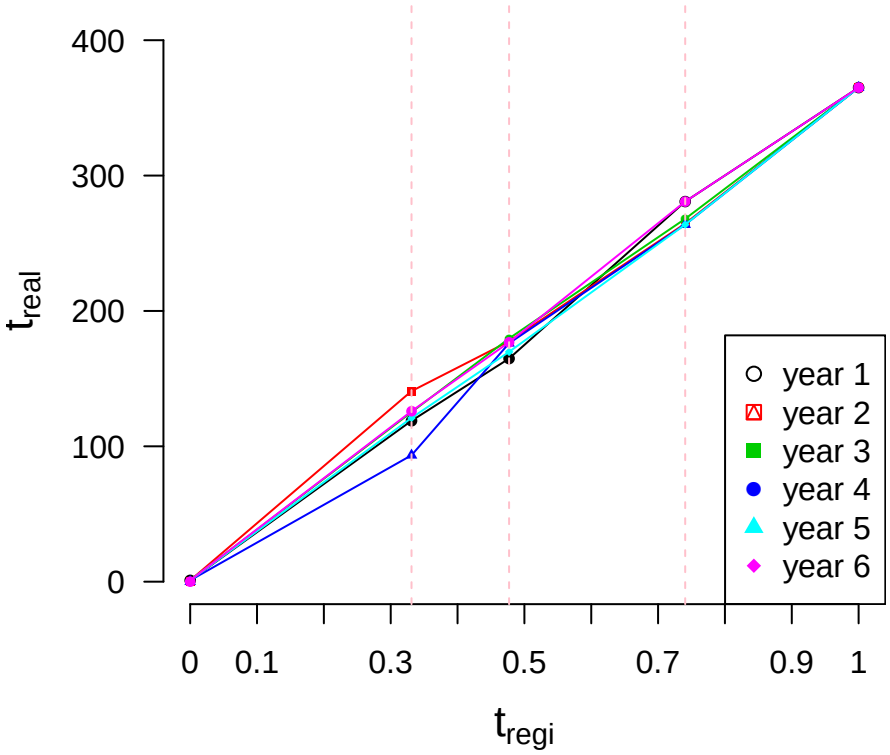


Figure 2.7: Warping function for old aspen. The vertical dashed lines indicate the three landmarks in the registered time scale. For example, the first landmark in the registered landmark is near 0.3.

For the old aspen data, we picked three landmarks every year, two at the peaks (around May and October), and one at the low point (near June). The reasons that we chose three landmarks are based on the pattern of the smoothed data and the seasonal cycle in North America. We use R function `locator` to select the landmarks for each year. The corresponding warping function is shown in Figure 2.7, whose y-axis is the time on the real time scale wherein the first landmarks have more variation than the other two, and the x-axis represents time on the registered time scale wherein the three vertical dashed lines indicate the location of the registered

landmarks t_l^*. We calculated t_l^* using

$$t_l^* = \frac{1}{365n} \sum_{i=1}^{n} t_{i,l}, \quad l = 1, \ldots, L. \tag{2.12}$$

Observations in the Registered Time Scale

Once the warping function is constructed, we know that for any given time on the real time scale, the corresponding time on the registered time scale is

$$t^* = h^{-1}(t). \tag{2.13}$$

This completes our introduction of the preliminary knowledge. The next Chapter introduces the overview of our single location modeling.

Chapter 3

Overview of the Single Location Modeling

The primary focus of this thesis is modeling the CO_2 flux data. This chapter describes the model for a single location, and Chapter 5 introduces how we handle multiple sites.

Section 3.1 displays the model overview and parameter estimation. Section 3.2 introduces the random sources and their connection to the single location modeling. More modeling details are elaborated in Chapter 4.

3.1 Overview of the Model

1. We assume that the annual CO_2 data f_i in year i follows a continuous process. At time t, the CO_2 flux will be $f_i(t)$. We assume f_i is composed of f and η_i, where f is the underlying and unobservable mean of all years of data and η_i is a continuous year effect.

$$f_i = f + \eta_i,\ i = 1, 2, \ldots, n, \tag{3.1}$$

where η_i is identically distributed with mean 0 and variance Σ_η, i.e $\Sigma_\eta(s,t) = Cov(\eta_i(s), \eta_i(t))$. We represent f_i using a linear combination of basis functions:

$$f_i(t) = \sum_{d=1}^{s+k+1} \beta_d \phi_d(t). \tag{3.2}$$

In matrix format, it is

$$f(t, \boldsymbol{\beta}) = \Phi(t)\boldsymbol{\beta}, \quad t \in [0, 365], \tag{3.3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{s+k+1})^T$ and $\Phi(t) = (\phi_1(t), \phi_2(t), \ldots, \phi_{s+k+1}(t))$ is a row vector. $\Phi(t)$ is similar to the definition of Φ_i in (2.8).

2. Because f_is have a similar annual pattern, but at shifted times each year, their underlying pattern is better captured with data registration. Since the model involves two time scales, we use asterisk ($*$) to denote the registered time scale. Thus, t and f_i concern real time; t^* and f_i^{**} are in the registered time scale. As previously mentioned, the range for t^* is arbitrary, and we use $[0, 1]$ for convenience.

   For time points $t^* = \frac{1}{365}, \frac{17}{365}, \ldots, \frac{353}{365}$, f_i and f_i^{**} have the following connection:

$$t = h_i(t^*), \tag{3.4}$$

$$f_i^{**}(t^*) = f_i(t), \tag{3.5}$$

   where h_i is the warping function for year i (see section 2.2.3).

   We model the registered process $f^*(t^*)$ through the 23 $f_i^{**}(t^*)$ points using basis function $\phi_d^*(t^*)$ and basis matrix Φ^*, whose definition is similar to (2.8).

$$f_i^*(t^*) = \sum \beta_d^* \phi_d^*(t^*), \tag{3.6}$$

$$= \Phi^* \boldsymbol{\beta}^*. \tag{3.7}$$

   The warping function h_i is modeled by landmarks. On the real time scale we assume the lengths of adjacent landmarks follow a rescaled *Dirichlet*($\boldsymbol{\alpha}$) distribution. On the registered time scale the landmark locations t_l^* can be chosen arbitrarily. In practice, we use the average of observed landmarks and linearly rescale them (2.12).

3. The observations y_{ij} are discrete, and they consist the discrete points of smoothed annual

data $f_i(t_j)$ and the measurement error ϵ_{ij}.

$$y_{ij} = f_i(t_j) + \epsilon_{ij} \tag{3.8}$$

$$= f(t_j) + \eta_i(t_j) + \epsilon_{ij}, \tag{3.9}$$

where $\epsilon_{ij} \overset{independent}{\sim} N(0, \sigma^2(t_j))$, $i = 1, 2, \ldots, n$, and $j = 1, 17, 33, \ldots, 353$.

The model can be briefly summarized as

$$y_{ij} \underset{\text{add } \epsilon_{ij}}{\overset{\text{smoothing}}{\rightleftharpoons}} f_i = f + \eta_i \underset{h_i^{-1}(t)}{\overset{h_i(t^*)}{\rightleftharpoons}} f_i^* = f + \eta_i^*. \tag{3.10}$$

Once we collect the discrete annual data y_{ij}, we can estimate the model parameters $(f, \Sigma_\eta, \boldsymbol{\alpha}, \Sigma_\epsilon)$ as follows:

- The underlying overall mean f.

  We first apply cubic spline smoothing with the same basis Φ_f independently to each year of observed data y_{ij} to estimate f_i,

  $$\hat{f}_i = \mathcal{S}(y_{ij}, \Phi_f), \tag{3.11}$$

  where $\mathcal{S}$ is the smoother.

  Then, the unobservable underlying process f is estimated using

  $$\hat{f} = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_i, \tag{3.12}$$

- The variance of the year effect $\Sigma_\eta(s, s)$ is estimated using

  $$\hat{\Sigma}_\eta(s, s) = \mathcal{S}((\hat{f}_i(s) - \hat{f}(s))^2, \Phi_\eta). \tag{3.13}$$

- The rescaled Dirichlet distribution parameter $\boldsymbol{\alpha}$.

  We model the length of adjacent landmarks $\ell_{i,b}$ (3.14) on the real time scale using a

rescaled Dirichlet distribution. Because the support for Dirichlet distribution is $[0, 1]$, but $\ell_{i,b} \in [0, 365]$, we rescale the modeling results by a factor of 365.

$$\ell_{i,b} = \begin{cases} t_{i,b} - 0, \text{ when } b = 1 \\ 365 - t_{i,b}, \text{ when } b = L + 1 \\ t_{i,b} - t_{i,b-1}, \text{ otherwise} \end{cases} \tag{3.14}$$

- For the variation of measurement noise Σ_ϵ, we calculate the residuals of year i day j first using:

$$e_{ij} = y_{ij} - \hat{f}_i(j). \tag{3.15}$$

We assume the measurement errors are independent. The variation of measurement error follows a continuous process, and we use e_{ij} to estimate it:

$$\hat{\Sigma}_\epsilon = \mathcal{S}(e_{ij}^2, \Phi_\epsilon). \tag{3.16}$$

3.2 Random Structure of the Model

Let's look at our model (3.10) again. The randomness comes from three sources:

1. the landmarks, which are used to model the warping function $h_i(t^*)$;
2. Dirichlet regression, which is used to construct the inverse warping function $h_i^{-1}(t)$;
3. the measurement error ϵ_{ij}.

We detail each random structure as follows:

(i) The warping function h_i is a strictly increasing function from the registered time scale to real time scale.

$$h_i : [0, 1] \mapsto [0, 365] \ \forall i. \tag{3.17}$$

The warping function h_i maps $(0, t_{i1}^*, t_{i2}^*, \ldots, t_{iL}^*, 1) \mapsto (0, t_{i1}, t_{i2}, \ldots, t_{iL}, 365)$, and is determined by linear interpolation in between.

(ii) The process f_i is obtained using

$$f_i(t) = f_i^*(h_i^{-1}(t)), \tag{3.18}$$

where $h_i^{-1}(t)$ is constructed by modeling the landmarks in real time scale using rescaled Dirichlet distribution:

$$(t_1, t_2, \ldots, t_L) \equiv D \sim \text{RDirichlet}(\alpha, 0, 365). \tag{3.19}$$

Therefore, (3.18) can be expressed as

$$f_i(t; \beta, D) = f_i^*(h_{i\,[\beta,D]}^{-1}(t), \beta). \tag{3.20}$$

(iii) The observations are modeled as

$$y_{ij} = f_i(t_j) + \epsilon_{ij}, \tag{3.21}$$

where f_i represents the idealized conditional mean for a random year i with registration and warping randomness (β, D). For the measurement noise, we assume that $\epsilon_{ij} \sim N(0, \sigma^2(t_j))$.

Chapter 4

Methodology of Single Location Modeling

In Chapter 2, we discussed the limitations in analyzing the data when treating it as a single time series. After noticing a similar underlying pattern starting at different times between May and June, we used spline smoothing treating each year independently. The potential of FDA landmark registration may help us to overcome the phase variation problem.

Chapter 3 introduced the single location model structure. The CO_2 flux data is modeled as observational error on a smooth curve f_i, where f_i is modeled as a linear combination of the cubic spline basis functions with coefficients $\boldsymbol{\beta}$. Then, we register f_i using the warping function $h_i(t^*)$ and denote the registered data as f_i^*. Next, we complete the inverse warping function $h^{-1}(t)$ by modeling the rescaled landmarks using a Dirichlet distribution. We then apply the inverse warping function to map the registered data back to the real time scale. Last, we add measurement error ϵ_{ij} on top of the previous steps. This model can be briefly summarized as $y_{ij} = f_i^*(h^{-1}_{i\,[\boldsymbol{\beta},D]}(t), \boldsymbol{\beta}) + \epsilon_{ij}$, where D stands for Dirichlet distribution.

This chapter details each modeling step. As mentioned in Section 2.2.3, proper registration can help reduce the phase variation. Beyond what we have introduced about registration, we want to add constraints, which guarantee the registered curves also have extreme values at the registered landmarks. Section 4.1 presents the registration with additional constraints. Section

4.2 describes the use of the multivariate normal (MVN) distribution to simulate random registered curves $f^*_{\mathtt{sim},i}(t^*)$. Section 4.3 talks about how we transform the simulated curves back to a real time scale using the inverse warping function, which requires that we model landmarks. To accomplish it, we use two approaches - Dirichlet regression and compositional data analysis (CDA). Because the fitting of Dirichlet regression is more interpretable, we choose Dirichlet regression to complete the inverse warping function $h_i^{-1}(t)$. After establishing $h_i^{-1}(t)$, the random registered curves are mapped back to the real time scale. In this way, we simulate the smoothed curves on the real time scale $f_{sim,i}(t)$. Section 4.4 discusses the modeling of measurement error ϵ_{ij}. Once we add ϵ_{ij} to $f_{sim,i}(t)$, we produce the simulated data $\tilde{y}_{ij}$. Now the modeling steps are complete. The quantities used to qualify the prediction uncertainty can be theoretically calculated, but this is not easy in practice. Instead, in Section 4.5 we use Monte Carlo simulation to approximate the distribution and study its properties. Section 4.6 presents a brief discussion, especially on the necessity of certain modeling steps.

Since the single location model has a hierarchical structure, we combine the section number and the process of modeling in Table 4.1.

Table 4.1: Steps in the modeling process.

Real Time		**Real Time**		**Registered Time**
y_{ij}	$\xrightarrow[\text{Section 2.2.3}]{\text{smoothing}}$	$\hat{f}_i(t)$	$\xrightarrow[\text{Sections 2.2.3 \& 4.1}]{\text{landmark registration}}$	$\hat{f}^*_i(t^*)$
				Section 4.2 $\downarrow$ simulation
$\tilde{y}_{ij}$	$\xleftarrow[\text{Section 4.4}]{\text{add } \epsilon_{ij}}$	$f_{simu,i}(t)$	$\xleftarrow[\text{Section 4.3}]{\text{complete } h_i^{-1}(t)}$	$f^*_{simu,i}(t^*)$

The datasets used in this chapter are:

1. For an example of single location modeling, we use SK-Old Aspen. The CO_2 flux data of SK-Old Aspen covered from 2003 to 2008, where 23 data points were collected each year. This dataset is used in Sections 4.1 to 4.4.
2. To compare Dirichlet regression and CDA, we want to test these two methods on multiple locations with similar environmental conditions. To better analyze the effect of latitude

and longitude, we picked 15 locations (3×5 grid), where for each location we collect the NDVI data from 2000 to 2014. This dataset is used in Section 4.3.

3. In Section 4.5, for the model checking purpose, we use 55 AmeriFlux sites, all of which have complete CO_2 flux data for at least five years.

4.1 Modeling the Registered Curves under Constraints

We determined the registered landmarks as $t_l^*, l = 1, 2, \ldots, L$ by equation (2.12). After specifying the landmarks, we need to constrain our fitted curves $f^*(t^*)$ to meet the conditions that define the landmarks. That is, we would like to fit the curve

$$f^*(t^*) = \sum_{d=1}^{s+k+1} \beta_d^* \phi_d^*(t^*), \quad t^* \in [0, 1], \tag{4.1}$$

with the following constraints

$$\begin{cases} f^{*\prime}(t_1^*) = \sum \beta_d^* \phi^{*\prime}{}_d(t_1^*) = 0 \\ f^{*\prime}(t_2^*) = \sum \beta_d^* \phi^{*\prime}{}_d(t_2^*) = 0 \\ \vdots \\ f^{*\prime}(t_L^*) = \sum \beta_d^* \phi^{*\prime}{}_d(t_L^*) = 0 \end{cases}, \tag{4.2}$$

i.e. the first derivatives of the curve $f^*(t^*)$ at the L landmarks are 0 on the registered time scale. Usually the number of landmarks is less than the number of basis functions, namely $L < s + k + 1$. In our case, $L = 3$ and $s + k + 1 = 13$. The details of solving the coefficients with constraints are as follows.

We present equation set (4.2) as its matrix form

$$\mathcal{A}_{L\times(s+k+1)} \boldsymbol{\beta}^*_{(s+k+1)\times 1} = \mathbf{0}, \tag{4.3}$$

where $\mathcal{A}_{lj} = \phi'_j(t^*_l)$. We then partition the $\mathcal{A}$ matrix into two matrices $(\mathcal{A}_{(1)}, \mathcal{A}_{(2)})$, where matrix $\mathcal{A}_{(2)}$ is an invertible L by L square matrix. This may require re-ordering the columns of $\mathcal{A}$. We compute the corresponding partition of $\boldsymbol{\beta}^*$ as $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^*_{(1)}, \boldsymbol{\beta}^*_{(2)})^T$, giving

$$\mathcal{A}\boldsymbol{\beta}^* = (\mathcal{A}_{(1)}, \mathcal{A}_{(2)}) \begin{pmatrix} \boldsymbol{\beta}^*_{(1)} \\ \boldsymbol{\beta}^*_{(2)} \end{pmatrix} = \mathcal{A}_{(1)}\boldsymbol{\beta}^*_{(1)} + \mathcal{A}_{(2)}\boldsymbol{\beta}^*_{(2)} = \mathbf{0}. \tag{4.4}$$

In this form, $\boldsymbol{\beta}^*_{(2)}$ is determined from $\boldsymbol{\beta}^*_{(1)}$ as

$$\boldsymbol{\beta}^*_{(2)} = \mathcal{A}^{-1}_{(2)}(-\mathcal{A}_{(1)}\boldsymbol{\beta}^*_{(1)}) = -\mathcal{A}^{-1}_{(2)}\mathcal{A}_{(1)}\boldsymbol{\beta}^*_{(1)}. \tag{4.5}$$

Each year's observation of $\boldsymbol{y}$ in the matrix form is

$$\boldsymbol{y}^* = \Phi^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}. \tag{4.6}$$

We partition the design matrix Φ^*, the same way as we partition $\mathcal{A}$. Then (4.6) can be written as

$$\boldsymbol{y}^* = (\Phi^*_{(1)}, \Phi^*_{(2)}) \begin{pmatrix} \boldsymbol{\beta}^*_{(1)} \\ \boldsymbol{\beta}^*_{(2)} \end{pmatrix} + \boldsymbol{\epsilon}. \tag{4.7}$$

After substituting equation (4.5) into (4.7), we can solve for $\boldsymbol{\beta}^*_{(1)}$,

$$\begin{aligned} \boldsymbol{y}^* &= (\Phi^*_{(1)}, \Phi^*_{(2)}) \begin{pmatrix} \boldsymbol{\beta}^*_{(1)} \\ -\mathcal{A}^{-1}_{(2)}\mathcal{A}_{(1)}\boldsymbol{\beta}^*_{(1)} \end{pmatrix} + \boldsymbol{\epsilon} = \Phi^*_{(1)}\boldsymbol{\beta}^*_{(1)} - \Phi^*_{(2)}\mathcal{A}^{-1}_{(2)}\mathcal{A}_{(1)}\boldsymbol{\beta}^*_{(1)} + \boldsymbol{\epsilon} \\ &= (\Phi^*_{(1)} - \Phi^*_{(2)}\mathcal{A}^{-1}_{(2)}\mathcal{A}_{(1)})\boldsymbol{\beta}^*_{(1)} + \boldsymbol{\epsilon} = \Phi^*\boldsymbol{\beta}^*_{(1)} + \boldsymbol{\epsilon}. \end{aligned} \tag{4.8}$$

The new design matrix Φ^* is $(\Phi^*_{(1)} - \Phi^*_{(2)}\mathcal{A}^{-1}_{(2)}\mathcal{A}_{(1)})$. We can estimate $\boldsymbol{\beta}^*_{(1)}$ by least squares as

$$\hat{\boldsymbol{\beta}}^*_1 = (\Phi^{*T}\Phi^*)^{-1}\Phi^{*T}\boldsymbol{y}^*. \tag{4.9}$$

After calculating $\hat{\boldsymbol{\beta}}^*_{(1)}$, we obtain $\hat{\boldsymbol{\beta}}^*_{(2)}$ using equation (4.5). The estimated coefficients of $\boldsymbol{\beta}^*$

are

$$\hat{\boldsymbol{\beta}}^* = \begin{pmatrix} \hat{\boldsymbol{\beta}}^*_{(1)} \\ \hat{\boldsymbol{\beta}}^*_{(2)} \end{pmatrix}. \tag{4.10}$$

The estimated registered curve can be presented as

$$\hat{f}^*_i(t^*) = \sum_{d=1}^{s+k+1} \hat{\beta}^*_{id}\phi^*_j(t^*),\ t^* \in [0,1]. \tag{4.11}$$

In Figure 2.6 (b), we fit the curve on the registered time scale without applying constraints. Figure 4.1 shows the result of enforcing the constraints. The differences are that the first peak (around May) of year 2 and second peak (around October) of year 4 are less prominent in Figure 4.1, while the valley (around July) of year 4 is lower. We see the constraints do not have a large effect. Nevertheless, we recommend the method described above, in case there are other datasets where this is not true.

4.2 A Model of the Registered Curves: Multivariate Normal

In practice, it is hard to theoretically calculate the model properties, such as prediction uncertainty. By means of Monte Carlo simulation, we can approximate the distribution of predictions and study many model properties. According to (4.1), to simulate curves in the registered time scale, what we actually need is to simulate β^*.

As a rough approximation to the true distribution, we model β^* using a multivariate normal distribution, whose mean $\boldsymbol{\mu}$ is calculated using

$$\mu_d = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}^*_{id}, \tag{4.12}$$

where $d = 1, \ldots, s + k + 1$.

We denote $\hat{\beta}^*_d = (\hat{\beta}^*_{1d}, \hat{\beta}^*_{2d}, \ldots, \hat{\beta}^*_{nd})$. Each element of the variance covariance matrix Σ_{β^*} is

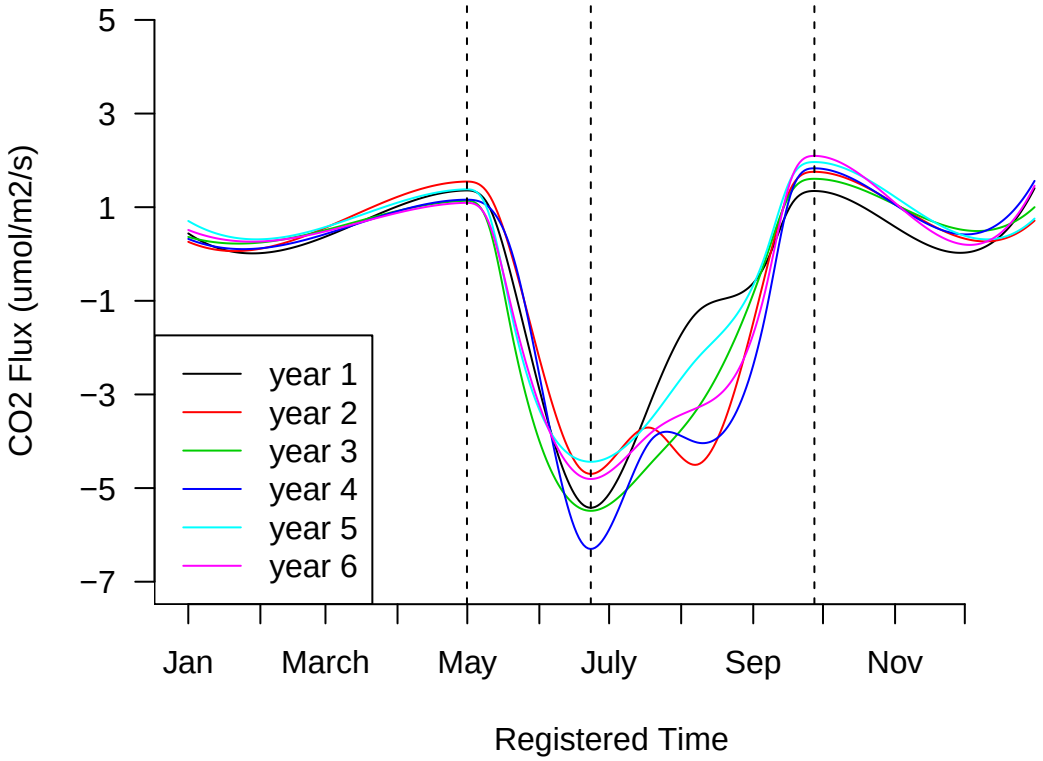


Figure 4.1: Under constraints registered curves of six-year old aspen data at three landmarks. The three vertical lines indicate the the location of three landmarks in the registered time scale.

calculated using

$$\sigma_{i,j} = cov(\hat{\beta}_i^*, \hat{\beta}_j^*). \tag{4.13}$$

Because the constraints (4.2) are linear, the sample covariance matrix will not be full rank and values simulated with that covariance will automatically obey the constraints.

The simulated data on the registered time scale is expressed as

$$f_{\texttt{sim}}^*(t^*) = \sum_{d=1}^{s+k+1} \beta_{\texttt{sim},d}^* \phi_d^*(t^*), \ t^* \in [0,1]. \tag{4.14}$$

We applied R function `mvrnorm` in package `MASS` (Venables and Ripley, 2002) to conduct

the simulation. We simulated 10,000 years of CO_2 flux data on the registered time scale. Figure 4.2 shows six simulations and the 95% point-wise prediction interval based on all 10,000 simulations. The registered Old Aspen data and the 95% point-wise prediction interval are shown in Figure 4.3.

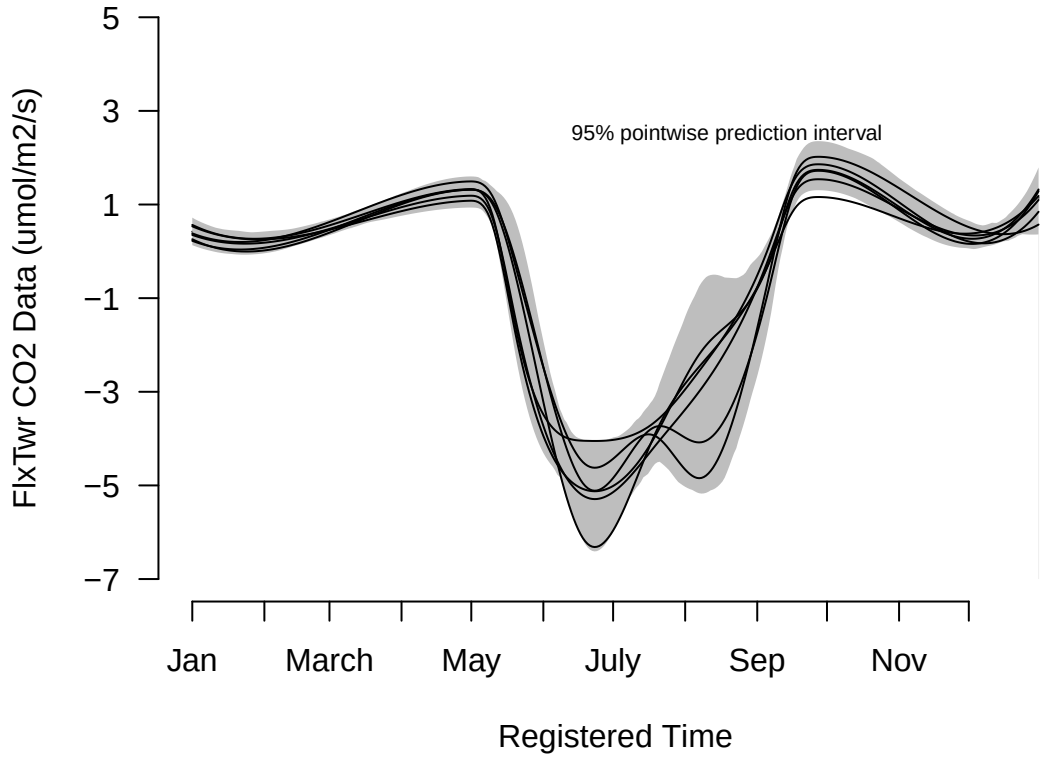


Figure 4.2: Six simulated curves of old aspen data on the registered time scale. The grey area is the 95% point-wise prediction interval.

4.3 A Model of the Inverse Warping Function

After simulating data in the registered time scale, we model the inverse warping function $h^{-1}(t)$ to map $f^*_{\texttt{sim}}(t^*)$ back to the real time scale for further study. Recall from our introduction to the warping function $h(t^*)$ introduced in Chapter 2, we first use landmarks in the real time

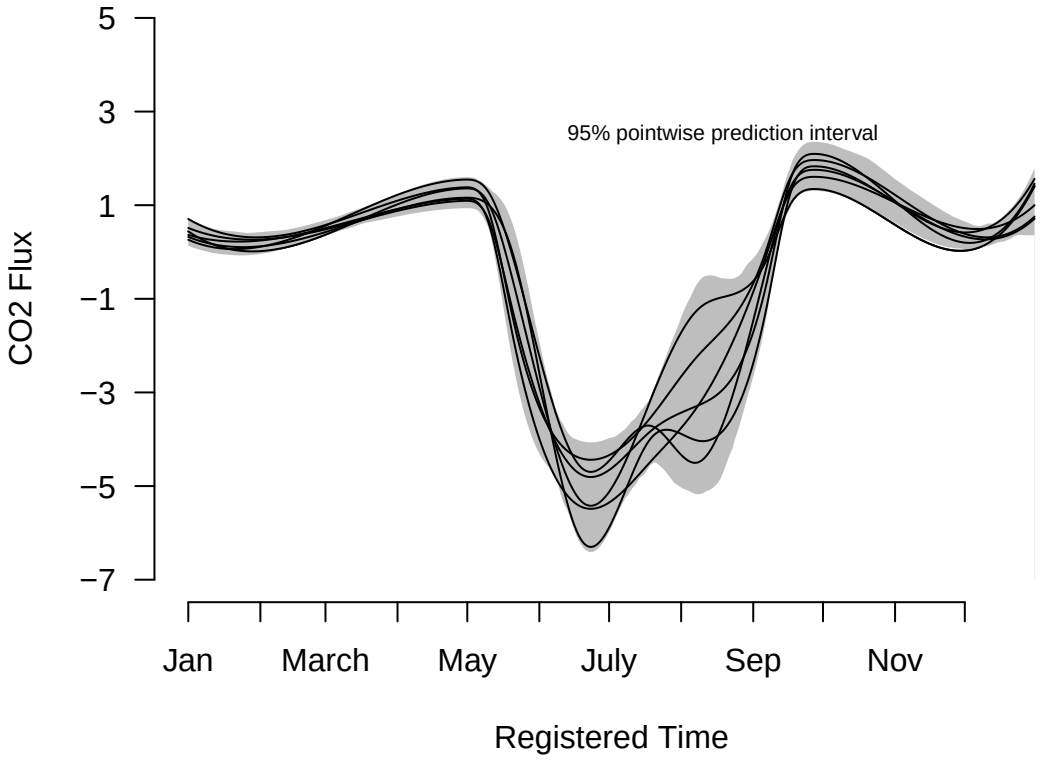


Figure 4.3: Six-year old aspen observed data on the registered time scale.

scale t_l to calculate the landmarks in the registered time scale t_l^*, then, we linearly interpolate them. For the inverse warping function, on the contrary, we have the landmarks in the registered time scale, and we need to model the landmarks on the real time scale to complete $h^{-1}(t)$. We cannot continue to use MVN to model the landmarks on the real time scale, because it would lead to the following problems:

1. range - we cannot guarantee that $1 \leq t_{sim,l} \leq 365, \ \forall l \in 1, 2, \ldots, L$;
2. order - we cannot guarantee that $t_{sim,1} < \cdots < t_{sim,L}$;
3. distance - we cannot avoid cases where $t_{sim,l}$ is very close to $t_{sim,l+1}$.

We might use rejection and sort to handle the problems 1 and 2, but there is no easy solution to the problem 3. Instead, we adopt different methods - Dirichlet regression and CDA, both of

which require us to consider the length of intervals between adjacent landmarks.

The i^{th} year's L unique landmarks divide the time interval $[0, 365]$ to the following $L + 1$ segments:

$$[0, t_{i,1}],\ [t_{i,1}, t_{i,2}],\ \ldots, [t_{i,L-1}, t_{i,L}], [t_{i,L}, 365]. \tag{4.15}$$

The length of the respective intervals are defined as:

$$\ell_{i,1} = t_{i,1},\ \ \ell_{i,2} = t_{i,2} - t_{i,1},\ \ldots,\ \ell_{i,L+1} = 365 - t_{i,L}, \tag{4.16}$$

and they satisfy the following properties:

1. $\ell_{i,b} > 0,\ \forall i, b = 1, 2, \ldots, L + 1$;
2. $\sum_{b=1}^{L+1} \ell_{i,b} = 365,\ \forall i$.

We rescale $\ell_{i,b}$ using

$$\tilde{\ell}_{i,b} = \frac{1}{365}\ell_{i,b},\ b = 1, \ldots, L + 1. \tag{4.17}$$

The rescaled intervals $\tilde{\ell}_{i,b}$ are all positive and summed up to 1.

4.3.1 Model Landmarks on t - Dirichlet Regression

The Dirichlet distribution $\mathcal{D}(\boldsymbol{\alpha})$ can describe the behaviour of the scaled intervals. It is a generalization of the Beta distribution. The Dirichlet distribution of order $z \geq 2$ with parameters $\alpha_1, \ldots, \alpha_z > 0$ has a probability density function given by

$$f(x_1, \ldots, x_z; \alpha_1, \ldots, \alpha_z) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{z} x_i^{\alpha_i - 1}, \tag{4.18}$$

where $\sum x_i = 1$, and $x_i > 0,\ \forall i = 1, \ldots, z$. The normalizing constant is the multivariate Beta function, which can be expressed in terms of the gamma function

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{z} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{z} \alpha_i)},\ \ \text{where } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_z). \tag{4.19}$$

The rescaled vector of intervals $\tilde{\boldsymbol{\ell}} = (\tilde{\ell}_1, \ldots, \tilde{\ell}_{L+1})$ follows a Dirichlet distribution $\mathcal{D}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{L+1})$. The sum of all parameters $\phi = \sum_{b=1}^{L+1} \alpha_b$ can be interpreted as precision. Under the same expected value, the higher the precision, the more concentrated the distribution. The expected value and variance for the b^{th} component are

$$E(\tilde{\ell}_b) = \mu_b = \frac{\alpha_b}{\phi}, \tag{4.20}$$

$$Var(\tilde{\ell}_b) = \sigma_b^2 = \frac{\alpha_b(\phi - \alpha_b)}{\phi^2(\phi + 1)}, \tag{4.21}$$

where $b = 1, \ldots, L + 1$.

Dirichlet Regression

Temperature and year are some potential covariates that may affect the behaviour of landmarks. We use $\boldsymbol{X}$ to present the collection of all potential covariates. The rescaled landmark segments $\tilde{\boldsymbol{\ell}}$ follow $\mathcal{D}(\boldsymbol{\alpha}(\boldsymbol{X}))$, where $\boldsymbol{\alpha}(\boldsymbol{X}) = (\alpha_1(\boldsymbol{X}), \ldots, \alpha_{L+1}(\boldsymbol{X}))$. Unlike other generalized linear models, we

Camargo et al. (2012) proposes each $\alpha_b(x_{i.})$, $b = 1, \ldots, L + 1$ to be a linear combination of $x_{i.}$:

$$\alpha_b(x_{i.}) = x_{i,1}\psi_{1,b} + x_{i,2}\psi_{2,b} + \cdots + x_{i,p}\psi_{p,b} = \boldsymbol{x}_{i.}\boldsymbol{\psi}_{.b}. \tag{4.22}$$

The goal is to estimate $\psi_{c,b}$, where $c = 1, \ldots, p$ and $b = 1, \ldots, L+1$ subject to the constraint $\alpha_b(x_{i.}) > 0$, $\forall b$. The parameters are calculated using maximum likelihood estimation (MLE). Under the assumption that $\tilde{\ell}_{1.}, \ldots, \tilde{\ell}_{n.}$ are independent, the likelihood function is

$$\mathcal{L}(\psi) = \prod_{i=1}^{n} \left[\Gamma(A_i(x_{i.})) \prod_{b=1}^{L+1} \frac{\tilde{\ell}_{ib}^{\alpha_b(x_{i.})-1}}{\Gamma(\alpha_b(x_{i.}))} \right], \tag{4.23}$$

where $A_i(x_{i.}) = \sum_{b=1}^{L+1} \alpha_b(x_{i.})$. The gradients of log-likelihood are

$$\frac{\partial \log \mathcal{L}(\psi)}{\partial \psi_{c,b}} = \sum_{i=1}^{n} \left\{ [\Psi(A_i(x_{i.})) - \Psi(\alpha_b(x_{i.})) + \log \tilde{\ell}_{ib}] x_{i,c} \right\}, \tag{4.24}$$

where $\Psi(u) = \frac{\partial log\Gamma(u)}{\partial u}$. The numerical solution used to compute the MLE requires to carefully choose the initial value, for which Hijazi and Jernigan (2009) propose a calculation method.

Maier (2014) uses a "GLM-like framework" for the parameter estimation in two different parameterizations.

1. The common parametrization

   The parameters are the same as mentioned in (4.18). The "GLM-like framework" works as below

   $$\log(\alpha_b(x_{i.})) = \boldsymbol{x}_{i.}\boldsymbol{\psi}_{.b}. \tag{4.25}$$

   Compared to (4.22), a log link function is applied to $\alpha_b(x_{i.}) > 0$.

2. The alternative parametrization

   In this case, the Dirichlet distribution is presented using its mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{L+1})$ and precision ϕ. Since $\mu_b \in (0, 1)$ and $\phi \in (0, +\infty)$, the logit and log link functions are convenient to use.

Maier (2015) wrapped the above implementation in an R package `DirichletReg`. We used it to simulate the scaled intervals.

We scale the $\tilde{\boldsymbol{\ell}}$ back to interval [0,365], then we can construct the inverse warping function using the simulated landmarks calculated using Dirichlet regression to generate the simulated CO_2 flux curves on the real time scale through

$$f_{sim}(t) = f^*_{sim}(t^*), \text{ where } t^* = h^{-1}(t). \tag{4.26}$$

4.3.2 Model Landmarks on t - CDA

Besides Dirichlet regression, compositional data analysis is an alternative approach to model the landmarks. Van den Boogaart and Tolosana-Delgado (2013) define a composition as a vector of R positive components summing up to a given constant. The lengths of the segments $\ell_{i,b}$ divided by landmarks calculated in (4.16) for any given year i fit the definition and can be

used as an example. We use $\boldsymbol{\mathcal{L}}$ to denote the n by $L+1$ matrix formed by all $\ell_{i,b}$, where $i = 1, n$ and $b = 1, \ldots, L+1$.

Briefly, the idea of CDA is to apply a transformation so that the compositional data is no longer restricted to a simplex. We can then model the transformed data using linear regression, and then apply an inverse transformation to coefficients and predictions. If we use $g(\cdot)$ to present a certain transformation, the model can be presented as

$$g(\boldsymbol{\mathcal{L}}) = \boldsymbol{X} g(\boldsymbol{\beta}_{CDA}). \tag{4.27}$$

The three commonly used transformations are

1. **Additive Log-ratio (alr) Transform**

$$alr(\boldsymbol{\ell}) = \left(\ln \frac{\ell_1}{\ell_{L+1}}, \ldots, \ln \frac{\ell_L}{\ell_{L+1}} \right) \tag{4.28}$$

   This method is potentially problematic because the distances between points in the transformed space are not the same for different divisors. (Pawlowsky-Glahn and Buccianti, 2011).

2. **Centered Log-ratio (clr) Transform**

   One way to avoid additive log-ratio transform's problem is to use the geometric mean as the divisor.

$$g(\boldsymbol{\ell}) = (\ell_1 \cdot \ell_2 \cdot \cdots \cdot \ell_{L+1})^{\frac{1}{L+1}} \tag{4.29}$$

$$clr(\boldsymbol{\ell}) = \left(\ln \frac{\ell_1}{g(\boldsymbol{\ell})}, \ldots, \ln \frac{\ell_{L+1}}{g(\boldsymbol{\ell})} \right) \tag{4.30}$$

   A disadvantage of this method is the covariance matrix of the *clr* transformed data is singular.

3. **Isometric Log-ratio (ilr) Transform**

Egozcue et al. (2003) proposed the isometric log-ratio transformation.

$$ilr(\boldsymbol{\ell}) = \ln(\boldsymbol{\ell})V, \tag{4.31}$$

where V is an $L+1$ by L Helmert Matrix.

$$V = \begin{bmatrix} \frac{L}{\sqrt{L(L+1)}} & 0 & \dots & 0 & 0 \\ \frac{-1}{\sqrt{L(L+1)}} & \frac{L-1}{\sqrt{L(L-1)}} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{-1}{\sqrt{L(L+1)}} & \frac{L-1}{\sqrt{L(L-1)}} & \dots & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{L(L+1)}} & \frac{L-1}{\sqrt{L(L-1)}} & \dots & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \tag{4.32}$$

Because the ilr transformation avoids the arbitrariness of *alr* and the singularity of *clr*, we chose it in our analysis.

4.3.3 Comparisons between Dirichlet Regression and CDA

We test Dirichlet regression and CDA on 15 locations. We want that all locations have similar ecosystems and all of them have at least ten years of data covering the same time period. However, the North America flux sites displayed did not provide sufficient data satisfying our expectation. Under this circumstance, we used MODIS data, which is available worldwide.

Because we already have some information about the area displayed in Fig 1.6, we chose 15 locations in this area and downloaded the NDVI data ranging from 2000 to 2014 using the R package `ModisTools`. To better analyze the latitude and longitude effects we chose the 15 locations on a 3×5 grid in Figure 4.4. The latitudes range from 53.6289°N to 54.0543°N, and the longitudes are from −106.1978°W to −104.7338°W. The red dots in Figure 4.4 indicate the 15 locations we selected, and the blue triangles are the flux sites we introduced in Chapter 1. Location number 11 coincides with SK Old Aspen.

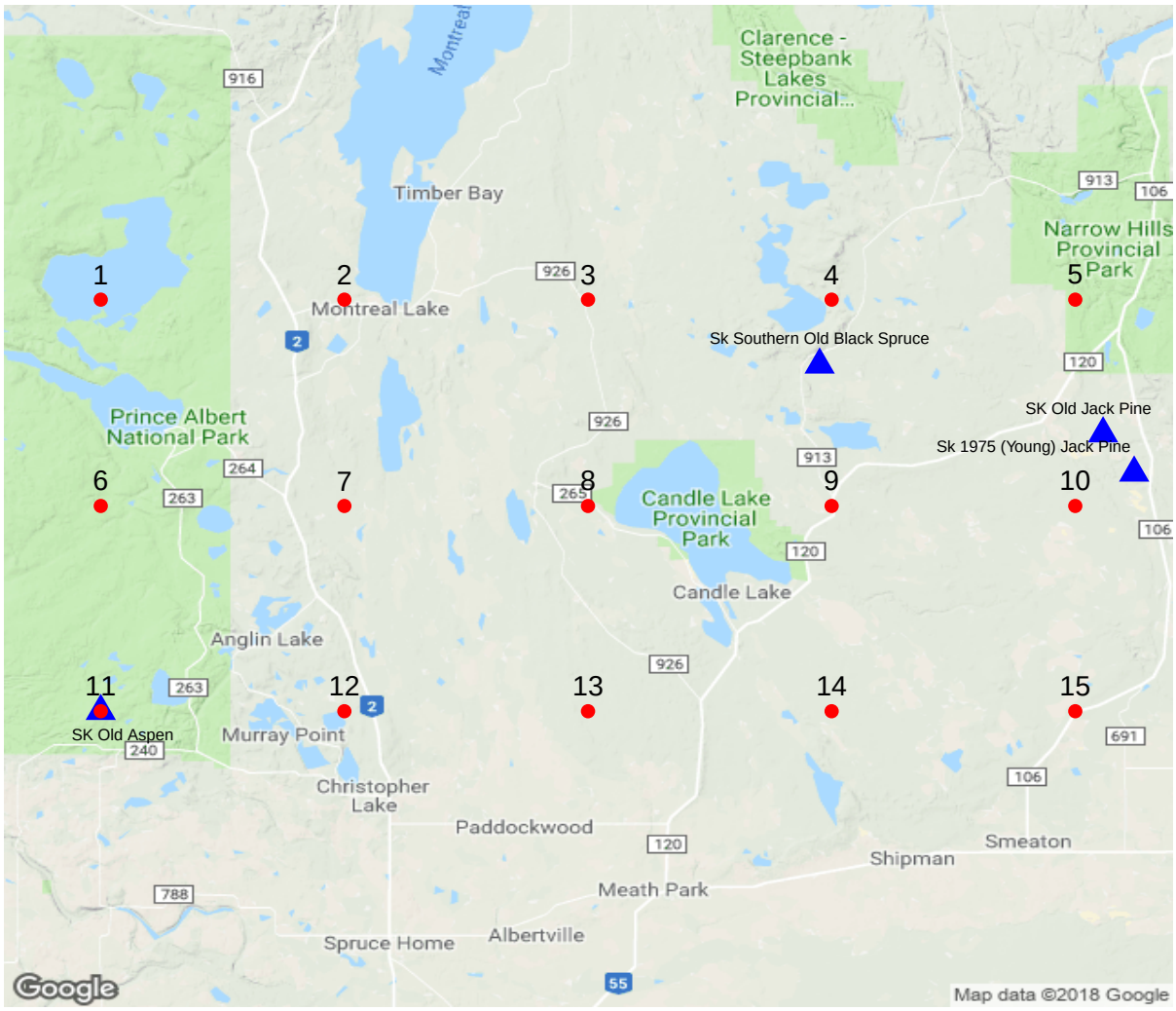


Figure 4.4: 15 locations chosen for testing Dirichlet regression and CDA.

Smoothed Curves for Each Location

The smoothed 15 curves of each site are plotted in Figure 4.5.

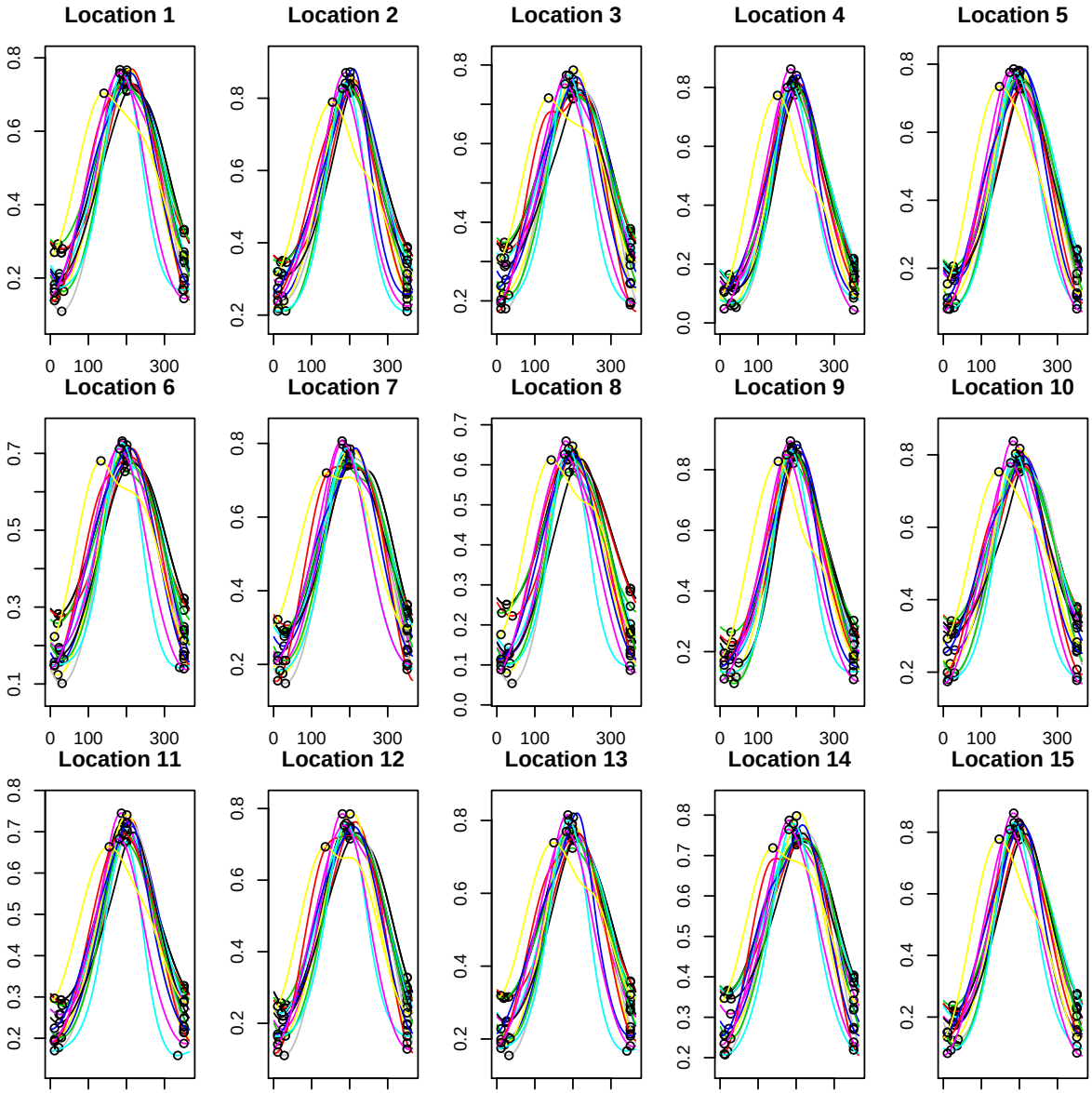


Figure 4.5: The smoothed 15 years of NDVI data from 2000 to 2014 for each location. We select three landmarks each year and circle them on the smoothed curves. The first and the third landmarks are close to the beginning and end of each year separately.

We use latitude, longitude and year as our covariates. For the 15 years of data we collected from each site, we set the last year as the testing dataset and the rest as the training dataset. We plot the predictions from CDA against predictions from Dirichlet regression in Figure 4.6. The dashed red line represents $y = x$ used as a reference. The predictions of the two methods for each segment are highly correlated.

We plot the observed landmarks and the predictions together in Figure 4.7. In the first row of Figure 4.7, we set the y limit of each graph to be identical. The predictions from the two methods are consistent. The predictions for segments 2 and 3 are less accurate than the predictions for segments 1 and 4. In the second row, the y limit are not the same, so we can see more details. Dirichlet regression predicts segments 1 and 4 larger than CDA.

For the same model, the compositional data analysis is faster than Dirichlet regression. Using a 1.3 GHz Intel Core m7 processor, the modeling time for Dirichlet regression is $4.057s$, but only $0.005s$ for CDA. This is because to compute the Dirichlet fitting we need a numeric solution to calculate MLE, while CDA uses matrix computation. Maier (2014) points out the drawbacks of CDA are the difficulty of interpreting the coefficients in the untransformed space, and inhomogeneity of the transformed data. Because in our study it is essential to interpret the coefficients in the context, because of (4.20) and (4.21), we choose Dirichlet regression. Figure 4.8 shows the effect of the inverse warping function produced by Dirichlet regression function on the results shown in Figure 4.2.

4.4 A Model of the Measurement Noise

In this section, we discuss the modeling process for the measurement error ϵ_{ij} in (3.8).

We assume ϵ_{ij} are independent but do not have constant variance and the vector $(\epsilon_1, \ldots, \epsilon_m)$

follows $MVN(0, \Sigma_\epsilon)$ with

$$\Sigma_\epsilon = \begin{bmatrix} \sigma^2(t_1) & 0 & \cdots & 0 \\ 0 & \sigma^2(t_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2(t_m) \end{bmatrix}. \tag{4.33}$$

The following steps describe how we model it:

1. We calculated the residuals between the observed data and the smoothed data in the real time scale using spline smoothing: $e_{ij} = y_{ij} - \hat{f}_i(t_j)$, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.
2. We applied spline smoothing to e_{ij}^2 with one interior knot point in the middle of the time interval $[0, 365]$ to estimate the variance $\sigma^2(t)$. The variance-covariance matrix of the measurement error is a diagonal matrix with $\sigma^2(t_j)$, $j = 1, 2, \ldots, m$ on the diagonal and zeros off the diagonal.

We use $\hat{\Sigma}_\epsilon$ to simulate measurement error. Figure 4.10 displays observations y_{ij} and simulations $\tilde{y}_{ij}$ of old aspen data.

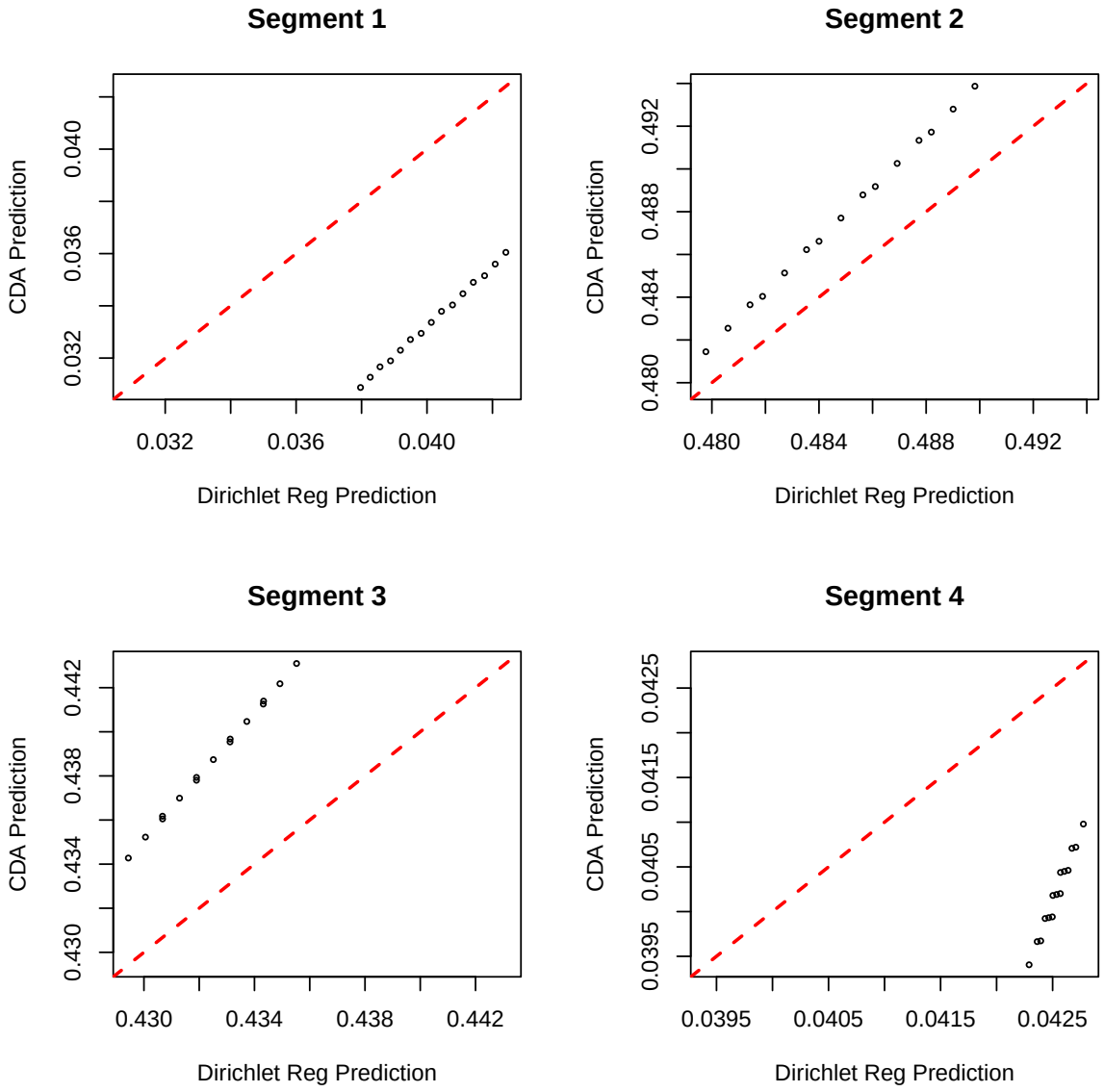


Figure 4.6: Prediction comparison between Dirichlet regression and compositional data analysis. The red dashed lines are lines $y = x$, which are used as the references. The black circles in each plot are highly linearly correlated. The predictions from the CDA for the second and third segments are larger than the predictions from Dirichlet regression. For the first and fourth segments, predictions from the Dirichlet regression are slightly bigger.

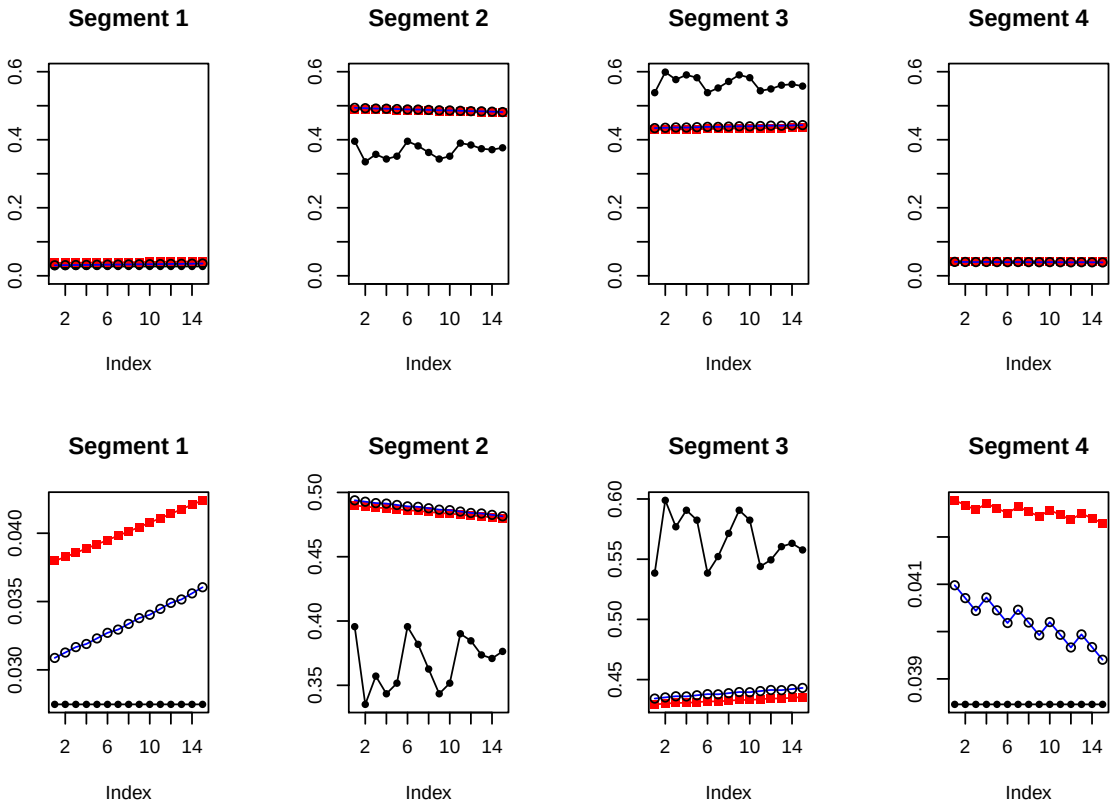


Figure 4.7: Landmark predictions and the observations. The black dots are the observations, the red squares are predictions from Dirichlet regression, and the blue circles are predictions from CDA. We plot the predictions of four segments with the same y limit in the first row. In this way, we can compare between segments and tell the overall trend. Segments 2 and 3 have larger effect than segments 1 and 2. The predictions for segment 2 are overall larger than the observations, while the predictions for segment 3 are smaller than the observed data. The predictions from the two method are similar. The second row plots predictions for each segment with adapted y limits. From these plots we can see the differences of predictions between two methods.

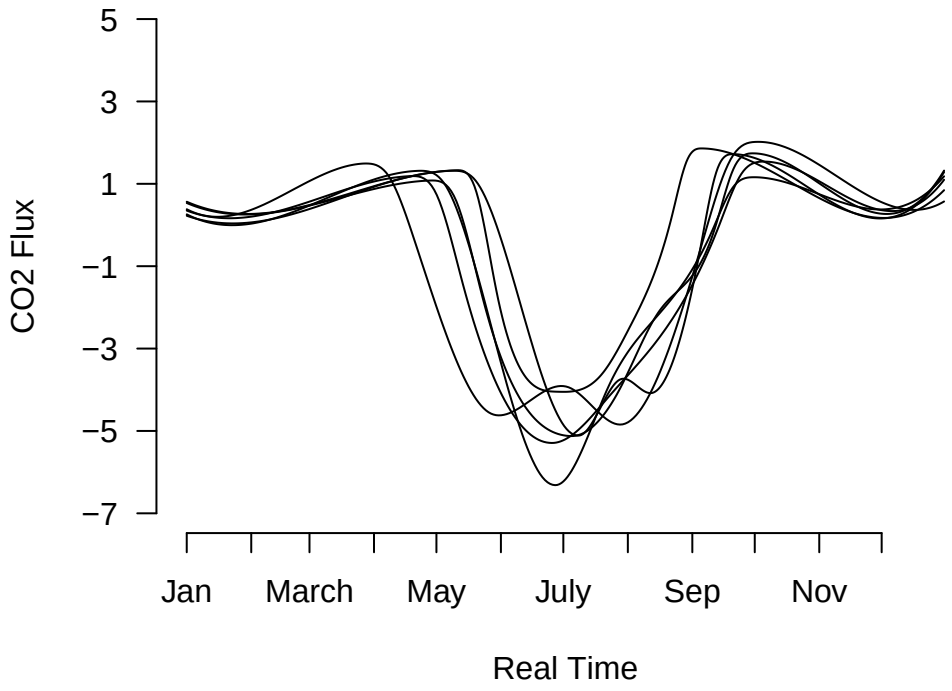


Figure 4.8: Six simulated curves of old aspen data on the real time scale based on Dirichlet fit. These are the results of inverse warped curves from Figure 4.3.

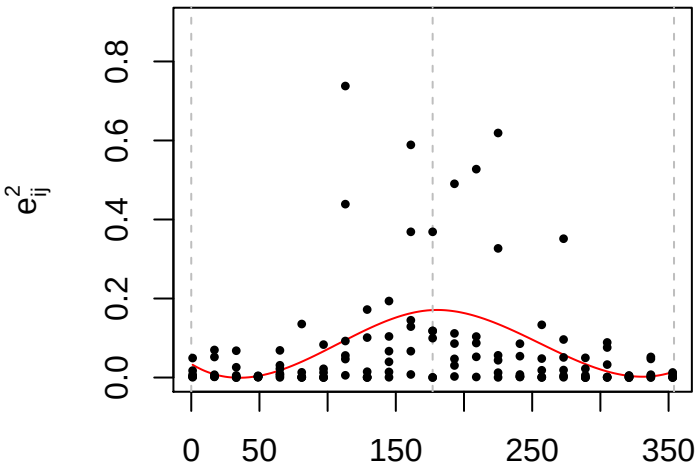


Figure 4.9: Old aspen measurement error variation estimation using equal length knots. The fitted curve has a larger value in the middle than on the two ends. The vertical dashed lines indicate the location we place knots.

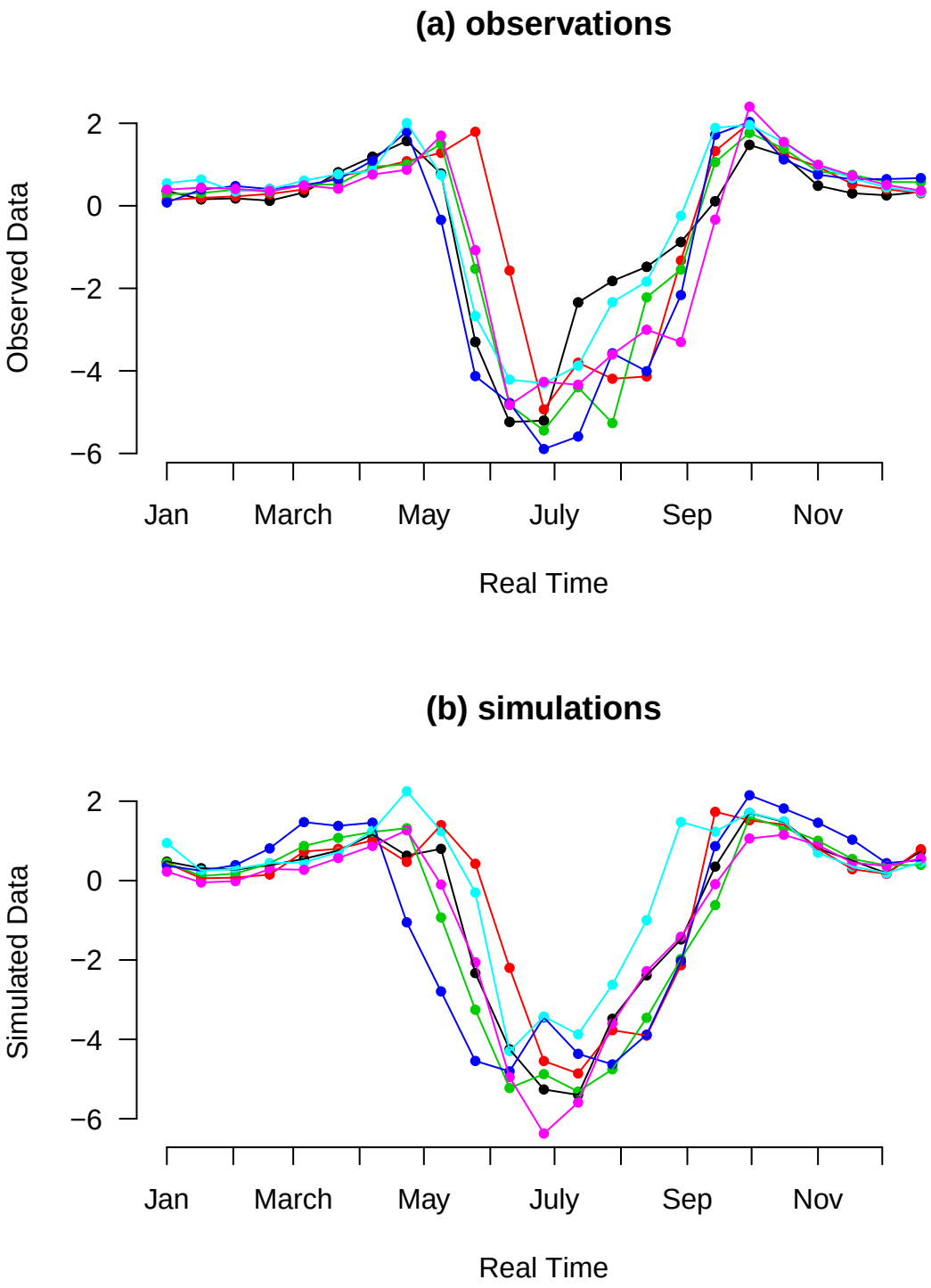


Figure 4.10: The observed and simulated data of old aspen. We plot the six-year observations in the top and the six-year simulations in the bottom. Each year has 23 data points. The simulations and observations share resemblance.

4.5 Model Checking

To assess the behavior of our model, we can use either mathematical derivation or simulation. Because our model has a hierarchical structure, simulation is much easier in practice. We chose the parametric bootstrap to construct the prediction interval for new observations. We evaluate the validity of our prediction interval by checking the rate at which simulations and observations fall in it. We applied this model checking method on 55 AmeriFlux sites, all of which have more than five years of complete data.

4.5.1 Parametric Bootstrap

Efron (1979) first introduced the non-parametric bootstrap. Suppose we have a sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and the statistic of interest is $G = g(X_1, X_2, \ldots, X_n)$. To perform his original bootstrap, there are four steps. First, we sample from our original sample $\boldsymbol{x}$ with replacement and obtain a new set of samples $\boldsymbol{x^*} = (x_1^*, x_2^*, \ldots, x_n^*)$. Then we compute one estimate of G based on the resampled data $\boldsymbol{x^*}$, i.e. $g_1 = g(x_1^*, x_2^*, \ldots, x_n^*)$. Next, we repeat the previous two steps for $N - 1$ times and generate $g_2, \ldots, g_N$. In the end, we use $(g_1, \ldots, g_N)$ to approximate the distribution of our statistic of interest.

In our project, we used the parametric bootstrap. Instead of sampling from the original dataset with replacement, samples are drawn from a fitted model. The following section explains the specifics of the parametric bootstrap step-by-step in the context of our study. For each selected site, we assume the original data set $\mathcal{M}_{n\times m}$ has n years of data and m data points each year, where y_{ij} denotes the observation of year $i = 1, \ldots, n$ at time $j = 1, \ldots, m$. We use $\boldsymbol{y}_i$ to denote the vector of data in year i.

1. For each site, we used the data in $\mathcal{M}$ to estimate all the parameters $\hat{f}, \hat{\Sigma}, \hat{\boldsymbol{\alpha}}$ and $\hat{\Sigma}_\epsilon$. In this way, we gain a fitted model. The details of estimation have been introduced in Section

3.1. To estimate a new observation $y_{i_0 j_0}$ at year i_0 time t_{j_0} we use:

$$\hat{y}_{i_0 j_0} = \hat{f}(t_{j_0}). \tag{4.34}$$

To assess how good the estimator $\hat{y}_{i_0 j_0}$ is, in addition to calculating one prediction of $y_{i_0 j_0}$, we also need to find the distribution of the difference between the estimator and the future value:

$$\delta_{j_0} = \hat{y}_{i_0 j_0} - y_{i_0 j_0} = \hat{f}(t_{j_0}) - y_{i_0 j_0}. \tag{4.35}$$

2. To achieve this, we use the parametric bootstrap. We generate new sets of samples from the fitted model, and obtain a set of $(\hat{y}^*_{i_0 j_0}, y^*_{i_0 j_0})$, also known as $(\hat{f}^*(t_{j_0}), y^*_{i_0 j_0})$. Here the asterisk sign denotes the quantity pertaining to the fitted model.

   We use the fitted model to simulate a set of data $\mathcal{M}^*$, whose elements are y^*_{ij}, $i = 1, 2, \ldots, n+1$ and $j = 1, 2, \ldots, m$. Because we treat each year $\boldsymbol{y}_i$ independently and identically, we set the last year of simulation $y^*_{(n+1) j_0}$ to be the new observation:

$$y^*_{i_0 j_0} = y^*_{(n+1) j_0}. \tag{4.36}$$

   We use the first n years of simulated data to estimate $\hat{f}^*$. According to (3.12), $\hat{f}^*$ is estimated using:

$$\hat{f}^* = \frac{1}{n} \sum_{i=1}^{n} f^*_i \tag{4.37}$$

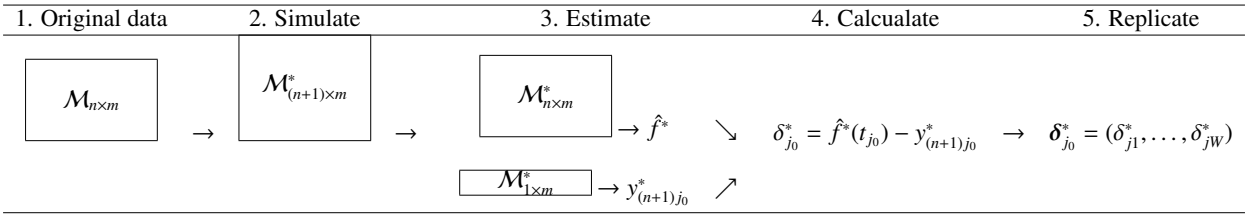
3. We calculate the difference between $\hat{f}^*(t_{j_0})$ and $y^*_{i_0 j_0}$:

$$\delta^*_{j_0} = \hat{f}^*(t_{j_0}) - y^*_{i_0 j_0} = \hat{f}^*(t_{j_0}) - y^*_{(n+1) j_0} \tag{4.38}$$

4. We repeat the previous two steps for W times, and gain $\boldsymbol{\delta}^*_{j_0} = (\delta^*_{j_0 1}, \ldots, \delta^*_{j_0 W})$. The distribution of δ_{j_0} (4.35) is approximated using the distribution of $\boldsymbol{\delta}^*_{j_0}$. Hence, the prediction interval of $y_{i_0 j_0}$ can be built. We illustrate the details of the prediction interval calculation

in the next section.

To summarize the steps of parametric bootstrap, we use the following chart:



4.5.2 Prediction Intervals

We applied the parametric bootstrap and produced $\boldsymbol{\delta}^*_{j_0}$, the distribution of which is an approximation of the distribution of $\hat{y}_{i_0 j_0} - y_{i_0 j_0}$. In this section, we focus on constructing local and global prediction intervals for $y_{i_0 j_0}$. Prediction intervals can be calculated for any required coverage level. We describe the process for 95% coverage.

- Local Prediction Interval

  A 95% point-wise prediction interval (q_1, q_2) satisfies:

$$P(q_1 \leq y_{i_0 j_0} \leq q_2) = 0.95. \tag{4.39}$$

  We chose the equal-tailed interval and calculated (q_1, q_2) using the following steps:

  1. For time point t_{j_0}, calculate the 2.5% and 97.5% quantiles $(L_{2.5\%,t_{j_0}}, U_{97.5\%,t_{j_0}})$ of $\boldsymbol{\delta}^*_{j_0}$.
  2. Since

$$\begin{aligned} 0.95 &= P(L_{2.5\%,t_{j_0}} \leq \delta_{j_0} \leq U_{97.5\%,t_{j_0}}) && (4.40)\\ &= P(L_{2.5\%,t_{j_0}} \leq \hat{y}_{i_0 j_0} - y_{i_0 j_0} \leq U_{97.5\%,t_{j_0}}) && (4.41)\\ &= P(L_{2.5\%,t_{j_0}} \leq \hat{f}(t_{j_0}) - y_{i_0 j_0} \leq U_{97.5\%,t_{j_0}}), && (4.42) \end{aligned}$$

the 95% point-wise prediction interval of $y_{i_0 j_0}$ is

$$[\hat{f}(t_{j_0}) - U_{97.5\%,t_{j_0}}, \hat{f}(t_{j_0}) - L_{2.5\%,t_{j_0}}]. \tag{4.43}$$

- Global Prediction Band

  Unlike the local prediction interval focusing only on a particular time point t_{j_0}, the global prediction interval applies to every time point t_j in a given year. The global prediction interval has a vector of lower limits $\boldsymbol{Q}_1 = (Q_{11}, Q_{12}, \ldots, Q_{1m})$ and a vector of upper limits $\boldsymbol{Q}_2 = (Q_{21}, Q_{22}, \ldots, Q_{2m})$. The global prediction band of $y_{i_0 j}$ is abbreviated as $(\boldsymbol{Q}_1, \boldsymbol{Q}_2)$, which satisfies:

$$P(Q_{1j} \leq y_{i_0 j} \leq Q_{2_j}, \forall t_j \text{ in year } i_0) = 0.95. \tag{4.44}$$

  There are many possible solutions to (4.44). We choose equal point-wise coverage at all time points. That is:

$$P(\alpha_1 \leq F_{t_j}(\hat{f}(t_j) - y_{i_0 j}) \leq \alpha_2, \forall t_j \text{ in year } i_0) = 0.95, \tag{4.45}$$

  where Q_i and α_i have the following relationship:

$$Q_{ij} = \hat{f}(t_j) - F^{-1}_{t_j}(\alpha_i), \quad i = 1, 2. \tag{4.46}$$

  The 95% global prediction band of $\boldsymbol{y}_{i_0}$ is made up of intervals

$$[\hat{f}(t_j) - F^{-1}_{t_j}(\alpha_1), \hat{f}(t_j) - F^{-1}_{t_j}(\alpha_2)], \tag{4.47}$$

  for $j = 1, \ldots, m$.

  We use $\boldsymbol{\Delta}^*$ to present the matrix of all δ^*_{jw}, where $j = 1, \ldots, m$ and $w = 1, \ldots, W$. Each row of $\boldsymbol{\Delta}^*$ is a vector $\boldsymbol{\delta}^*_j = (\delta^*_{j1}, \ldots, \delta^*_{jW})$. We calculate α_1 and α_2 using the following steps:

1. We calculate the percentile of $\boldsymbol{\delta}^*_j$ for $j = 1, \ldots, m$. The smallest number of $\boldsymbol{\delta}^*_j$ is mapped to 0, the largest is mapped to 1, and the rests are uniformly interpolated between 0 and 1.
2. We use $\boldsymbol{p}^*_j$ to denote the percentiles associated with $\boldsymbol{\delta}^*_j$. The percentile for δ^*_{jw} is denoted by p^*_{jw}, all of which forms a matrix of percentiles $\boldsymbol{P}^*$, whose row is presented using $\boldsymbol{p}^*_{j\cdot}$ and the column is denoted by $\boldsymbol{p}^*_{\cdot w}$. The range for every $\boldsymbol{p}^*_{j\cdot}$ is [0, 1], but the range for $\boldsymbol{p}^*_{\cdot w}$ varies from replication to replication .
3. For each replication of percentiles $\boldsymbol{p}^*_{\cdot w}$, we calculate its maximum and minimum. All the maximal percentiles form the vector $\boldsymbol{p}^*_{max} = (p^*_{max,1}, \ldots, p^*_{max,W})$, and all minimal percentiles produce the vector $\boldsymbol{p}^*_{min} = (p^*_{min,1}, \ldots, p^*_{min,W})$.
4. The α_1 is calculated by the 2.5% quantile of $\boldsymbol{p}^*_{min}$, *and* α_2 is calculated by the 97.5% quantile of $\boldsymbol{p}^*_{max}$.
5. After calculating α_1 and α_2, we calculate the 95% global prediction band of y_{i_0} using (4.47).

4.5.3 Coverage Rate

After the local and global prediction intervals are calculated, we can calculate their coverage from W simulations.

To calculate the local coverage rate of new observation $y^*_{i_0 j_0 w}$, we first define an indicator variable $\mathbb{1}^*_{local,j_0 w}$, which denotes whether the new observation at the w^{th} replicate at time t_{j_0} is in the interval (q_1, q_2):

$$\mathbb{1}^*_{local,j_0 w} = \begin{cases} 1, \ q_1 \leq y^*_{i_0 j_0 w} \leq q_2 \\ 0, \ otherwise \end{cases} . \tag{4.48}$$

The local coverage rate over all time points $t_{j_0} = 1, \ldots, m$ is calculated using:

$$\text{local rate} = \frac{\sum_{j_0=1}^{m} \sum_{w=1}^{W} \mathbb{1}^*_{local,j_0 w}}{m \times W}. \tag{4.49}$$

For the global coverage rate, we define an indicator variable $\mathbb{1}^*_{global,w}$ for $y^*_{i_0 jw}$. This indicator variable represents whether all the points of the w^{th} simulation are within the global prediction interval:

$$\mathbb{1}^*_{global,w} = \begin{cases} 1, \ \forall t_j, \ Q_{1j} \leq y^*_{i_0 jw} \leq Q_{2j}, \\ 0, \ otherwise \end{cases} . \tag{4.50}$$

The corresponding global coverage rate is:

$$\text{global rate} = \frac{\sum_{w=1}^{W} \mathbb{1}^*_{global,w}}{W}. \tag{4.51}$$

4.5.4 Coverage Results

We simulated on the 55 AmeriFlux sites first based on $W = 1,000$ replications but the global coverage rates were poorly estimated. The following results are based on $W = 10,000$.

Table 4.2 contains both global and local coverage rates in descending order of the number of years of data. The local and the global coverage rates all fluctuate around 95%.

Figure 4.11 and Figure 4.12 show there is no evidence that the number of years has a large effect on either the local or the global coverage rate. The group of sample size 5, which is the shortest among all sample sizes, has the largest variation in coverage. It's local coverage rate still seems to average about 95%, but the global coverage rate may be lower.

Figure 4.13 and Figure 4.14 display little evidence that the vegetation type affects the global coverage rate, even though the shapes of data of different vegetations are very different.

Table 4.2: Coverage rate by quantity of data.

	Sites	Years	LocalRate	GlobalRate	Vegetation
47	US-UMB	23	0.94	0.93	DBF
18	US-Ha1	20	0.95	0.94	DBF
20	US-Ho1	16	0.92	0.85	ENF
26	US-MMS	16	0.95	0.94	DBF
32	US-NR1	15	0.94	0.94	ENF
34	US-PFa	14	0.96	0.96	MF
49	US-Var	13	0.95	0.94	GRA
7	CA-TP4	12	0.95	0.96	ENF
44	US-SRM	12	0.95	0.95	WSA
29	US-Ne1	11	0.95	0.92	CRO
31	US-Ne3	11	0.95	0.95	CRO
45	US-Ton	11	0.95	0.95	WSA
54	US-Wkg	11	0.95	0.94	GRA
55	US-Wrc	11	0.96	0.96	ENF
27	US-MOz	10	0.94	0.92	DBF
30	US-Ne2	10	0.96	0.96	CRO
38	US-Slt	10	0.95	0.95	DBF
50	US-WBW	10	0.96	0.96	DBF
3	CA-Qcu	9	0.96	0.96	ENF
11	US-Bo1	9	0.95	0.95	CRO
17	US-GLE	9	0.94	0.91	ENF
21	US-Ho2	9	0.94	0.93	ENF
6	CA-TP3	8	0.95	0.96	ENF
8	US-ARM	8	0.94	0.94	CRO
14	US-Ced	8	0.95	0.95	CSH
15	US-Dk2	8	0.94	0.92	DBF
35	US-Ro1	8	0.96	0.96	CRO
37	US-Ses	8	0.95	0.96	OSH
52	US-Whs	8	0.94	0.92	OSH
1	BR-Sa1	7	0.96	0.96	EBF
16	US-Dk3	7	0.95	0.94	ENF
19	US-Ha2	7	0.95	0.95	ENF
23	US-IB2	7	0.95	0.96	GRA
25	US-Me2	7	0.95	0.95	ENF
28	US-Mpj	7	0.91	0.81	WSA
33	US-Oho	7	0.94	0.90	DBF
39	US-Snd	7	0.95	0.96	GRA
41	US-SP1	7	0.96	0.95	ENF
42	US-SP3	7	0.95	0.96	ENF
43	US-SRG	7	0.96	0.95	GRA
2	CA-Gro	6	0.95	0.95	MF
9	US-Bar	6	0.94	0.94	DBF
10	US-Blo	6	0.96	0.96	ENF
13	US-Br3	6	0.94	0.94	CRO
22	US-IB1	6	0.90	0.89	CRO
36	US-Seg	6	0.96	0.96	GRA
46	US-Twt	6	0.96	0.97	CRO
48	US-UMd	6	0.95	0.94	DBF
51	US-WCr	6	0.96	0.96	DBF
4	CA-Qfo	5	0.96	0.96	ENF
5	CA-TP1	5	0.96	0.96	ENF
12	US-Br1	5	0.96	0.96	CRO
24	US-Los	5	0.90	0.79	WET
40	US-SO2	5	0.93	0.79	CSH
53	US-Wjs	5	0.93	0.85	SAV

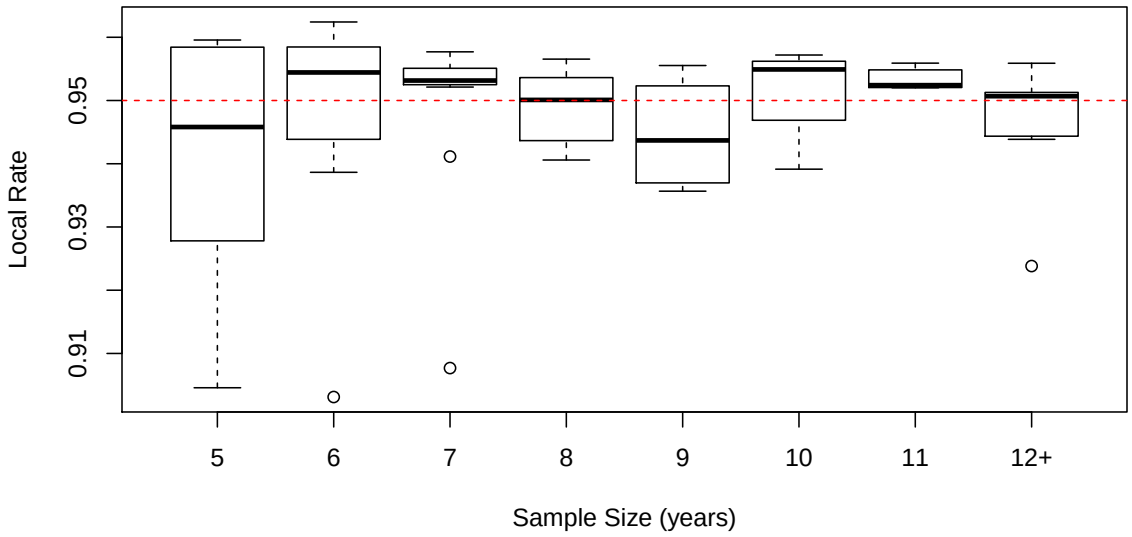


Figure 4.11: Local coverage rate grouped by the number of years. The median of the local coverage rate of each year group fluctuates around 0.95. There is no obvious trend that as the number of observed year increases the variations of local coverage rate decrease, but the group with sample size of 5 years, which is the smallest, has the largest variation in local coverage rate.

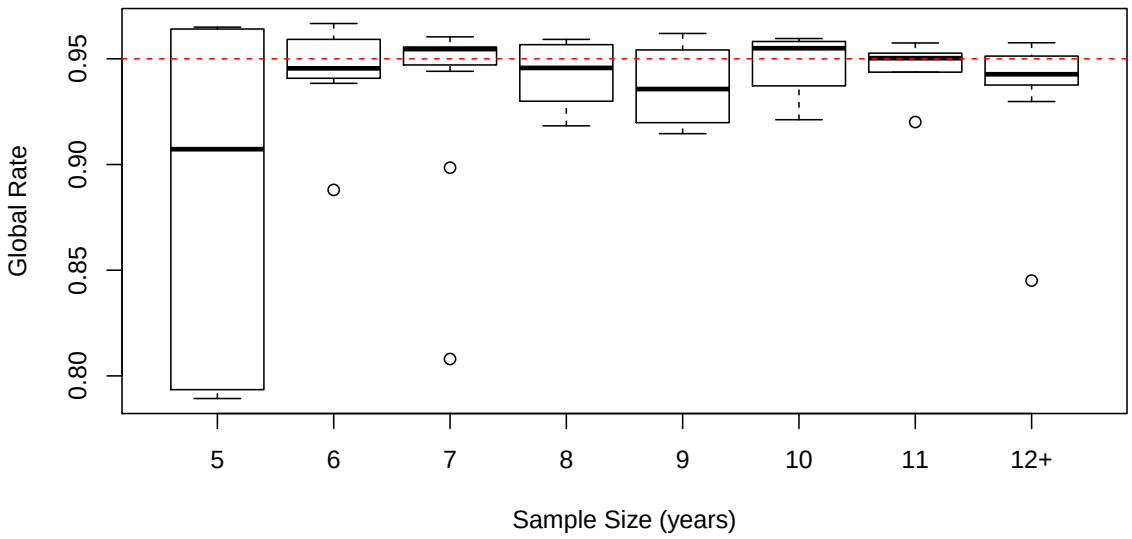


Figure 4.12: Global coverage rate grouped by the number of years. This image has similar pattern with Figure 4.11. The variation in group with sample size of 5 years is obviously larger than other groups, and the first quantile of this group is around 0.8, while the first quantiles of other groups are all above 0.9.

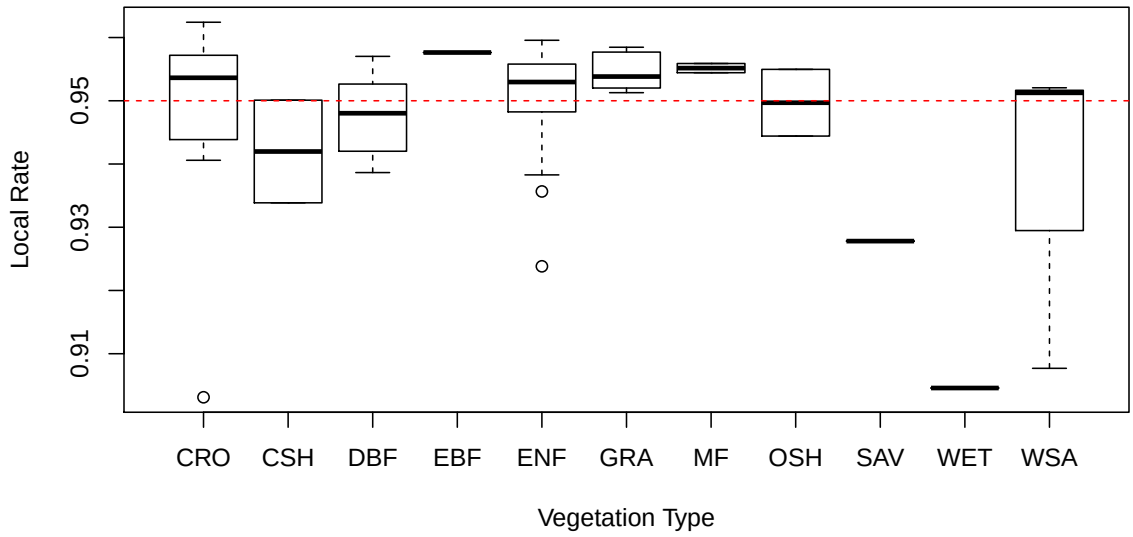


Figure 4.13: Local coverage rate grouped by the vegetation types. Among 11 vegetation types, four of them - CSH, SAV, WET and WSA, behave not ideally. The box plot of CSH barely covers the 0.95. For EBF, SAV and WET, the shapes of plots indicate these vegetation types may lack of data. As for WSA, the distribution is highly skewed.

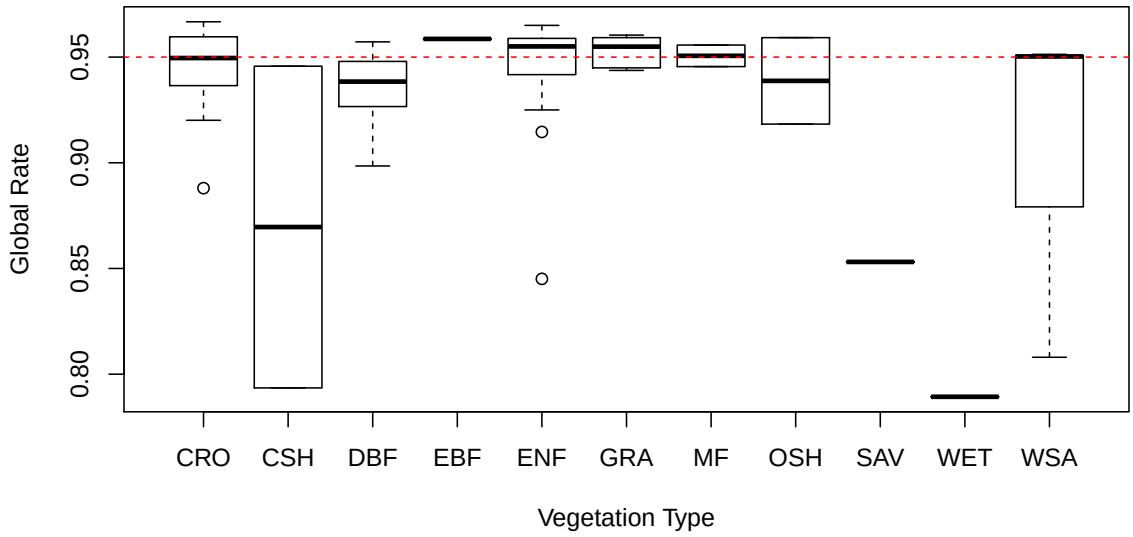


Figure 4.14: Global coverage rate grouped by the vegetation types. This image is similar to Figure 4.13. The medians of the global coverage rates of all vegetation types are generally lower than the local coverage rates. The variations in the global rates are overall larger.

4.5.5 Observations and Prediction Intervals

In this section, we explore the local and global coverage rates on our observations. Similar to what we explained in Section 4.5.3, the difference is that instead of checking whether $y^*_{i_0 j_0 w}, w = 1, \ldots, W$ inside the intervals, which is defined in (4.52) and (4.54), we check on the observations $y_{ij}, i = 1 \ldots, n$, and $j = 1, \ldots, m$ for each site.

The observations' local coverage rate is calculated using:

$$\mathbb{1}_{local,ij} = \begin{cases} 1, \ q_1 \leq y_{ij} \leq q_2 \\ 0, \ otherwise \end{cases}, \tag{4.52}$$

$$\text{observations' local rate} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{local,ij}}{n \times m}. \tag{4.53}$$

The observations' global coverage rate is calculated using:

$$\mathbb{1}_{global,i} = \begin{cases} 1, \ \forall t_i \ Q_{1j} \leq y_{ij} \leq Q_{2j} \\ 0, \ otherwise \end{cases}, \tag{4.54}$$

$$\text{observations' global rate} = \frac{\sum_{i=1}^{n} \mathbb{1}_{global,i}}{n}. \tag{4.55}$$

The results are shown in Table 4.3. For observations' local rates, the majority are above 95%. For the global rates, most of them are over 90%, but site US-SO2 is only 40% and the rate for US-SP1 is around 70%. To find the reasons, we plot the observations, local and global bands in the same plot for each site together with the smoothed data in Figure 4.15. Figure 4.16 shows two examples, sites US-Los and US-Me2, of how extreme values influence the widths of prediction bands. Figure 4.17 shows sites CA-Gro and CA-Qfo, both of which have good local and global rates.

Table 4.3: Each site observations' coverage rates.

	Site	Years	Local.Rate	Global.Rate	Vegetation
47	US-UMB	23	0.97	1.00	DBF
18	US-Ha1	20	0.98	0.85	DBF
20	US-Ho1	16	0.99	1.00	ENF
26	US-MMS	16	0.96	0.94	DBF
32	US-NR1	15	0.98	1.00	ENF
34	US-PFa	14	0.98	1.00	MF
49	US-Var	13	0.97	0.92	GRA
7	CA-TP4	12	0.99	1.00	ENF
44	US-SRM	12	0.98	0.92	WSA
29	US-Ne1	11	0.98	1.00	CRO
31	US-Ne3	11	0.99	1.00	CRO
45	US-Ton	11	0.97	0.91	WSA
54	US-Wkg	11	0.97	0.91	GRA
55	US-Wrc	11	0.98	0.91	ENF
27	US-MOz	10	1.00	1.00	DBF
30	US-Ne2	10	1.00	1.00	CRO
38	US-Slt	10	0.97	0.90	DBF
50	US-WBW	10	0.98	1.00	DBF
3	CA-Qcu	9	0.98	1.00	ENF
11	US-Bo1	9	0.98	1.00	CRO
17	US-GLE	9	0.99	1.00	ENF
21	US-Ho2	9	0.99	1.00	ENF
6	CA-TP3	8	0.99	1.00	ENF
8	US-ARM	8	0.99	0.88	CRO
14	US-Ced	8	0.97	1.00	CSH
15	US-Dk2	8	0.98	1.00	DBF
35	US-Ro1	8	0.98	0.88	CRO
37	US-Ses	8	0.97	1.00	OSH
52	US-Whs	8	0.98	1.00	OSH
1	BR-Sa1	7	0.96	0.86	EBF
16	US-Dk3	7	0.97	0.86	ENF
19	US-Ha2	7	0.96	1.00	ENF
23	US-IB2	7	0.99	0.86	GRA
25	US-Me2	7	0.98	0.86	ENF
28	US-Mpj	7	0.93	1.00	WSA
33	US-Oho	7	0.99	1.00	DBF
39	US-Snd	7	0.98	1.00	GRA
41	US-SP1	7	0.99	0.71	ENF
42	US-SP3	7	0.97	1.00	ENF
43	US-SRG	7	0.98	1.00	GRA
2	CA-Gro	6	0.99	1.00	MF
9	US-Bar	6	0.99	1.00	DBF
10	US-Blo	6	0.98	1.00	ENF
13	US-Br3	6	0.99	1.00	CRO
22	US-IB1	6	1.00	1.00	CRO
36	US-Seg	6	0.99	1.00	GRA
46	US-Twt	6	0.99	1.00	CRO
48	US-UMd	6	0.99	1.00	DBF
51	US-WCr	6	0.99	1.00	DBF
4	CA-Qfo	5	0.97	1.00	ENF
5	CA-TP1	5	1.00	1.00	ENF
12	US-Br1	5	0.97	1.00	CRO
24	US-Los	5	0.99	0.80	WET
40	US-SO2	5	0.92	0.40	CSH
53	US-Wjs	5	0.94	0.80	SAV

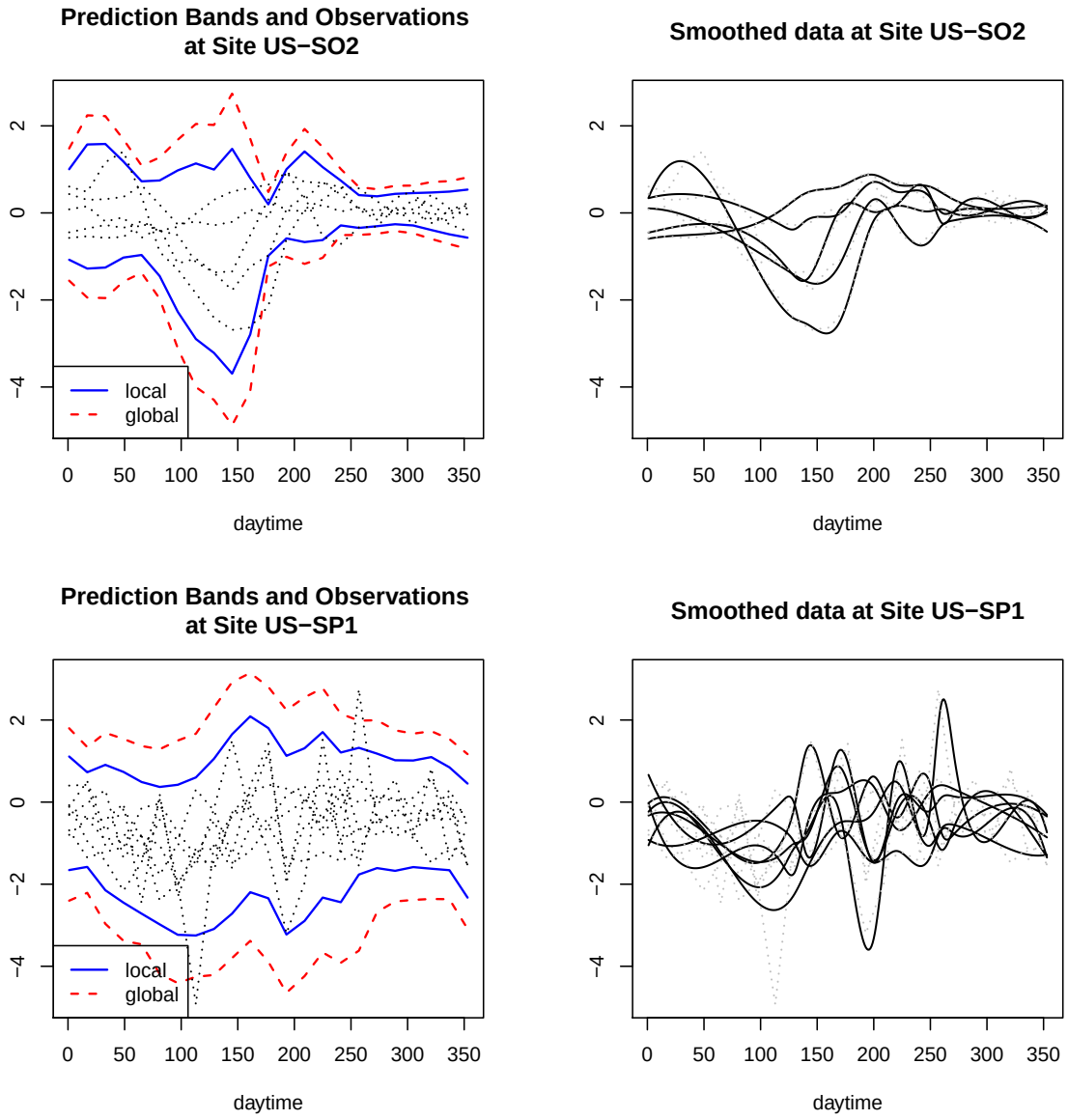


Figure 4.15: Observations' global coverage rate example 1. Three out of five years of US-SO2 data fail the global band. The smoothed data show that around day 40, day 150 and day 180 the smoothed observations have larger variation. The prediction bands capture the pattern. The global rate of US-SP1 using parametric bootstrap is 95%, but only 71% for observations. The reasons can be the smoothing fail to capture the extreme value around day 110, and the simulated data may not capture the extreme value around day 270.

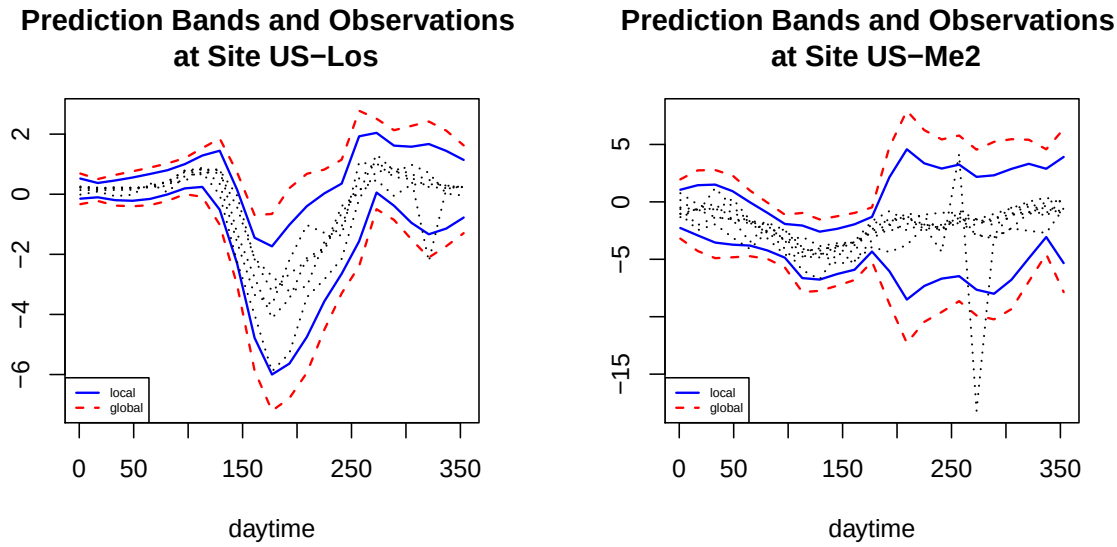


Figure 4.16: Observations' global coverage rate example 2. We use sites US-Los and US-Me2 to illustrate how extreme value affect the prediction bands. For site US-Los on the left around day 320, there is one observation has a sharp drop. Other than this point the observations at the beginning and the end of the year are with low variation. But due to one extreme point, both the local and global prediction bands from day 250 to day 365 are wider than the bands from day 0 to day 120. Another example is from site US-Me2 on the right. There are two extreme data points within day 200 to day 300. The prediction bands are inflated from day 150 to the end of the year.

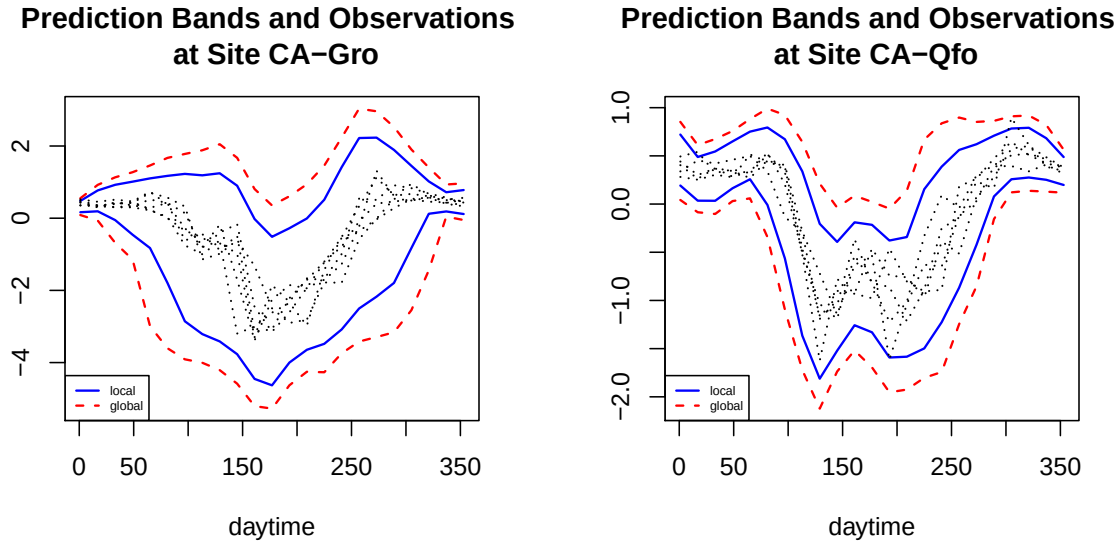


Figure 4.17: Observations' global coverage rate example 3. We use sites CA-Gro and CA-Qfo as a demonstration of good local and global prediction bands.

4.5.6 Vegetation Type and Prediction Intervals

For each site, we calculated the global prediction band $(\boldsymbol{Q}_1, \boldsymbol{Q}_2)$. The width of the global prediction interval is calculated using:

$$d(t_j) = Q_{2j} - Q_{1j}. \tag{4.56}$$

We ran the whole simulation twice and all the plots are pretty similar. We plot the bandwidths of four vegetation types in Figure 4.18. For better visual comparison, all the plots have the same y scale.

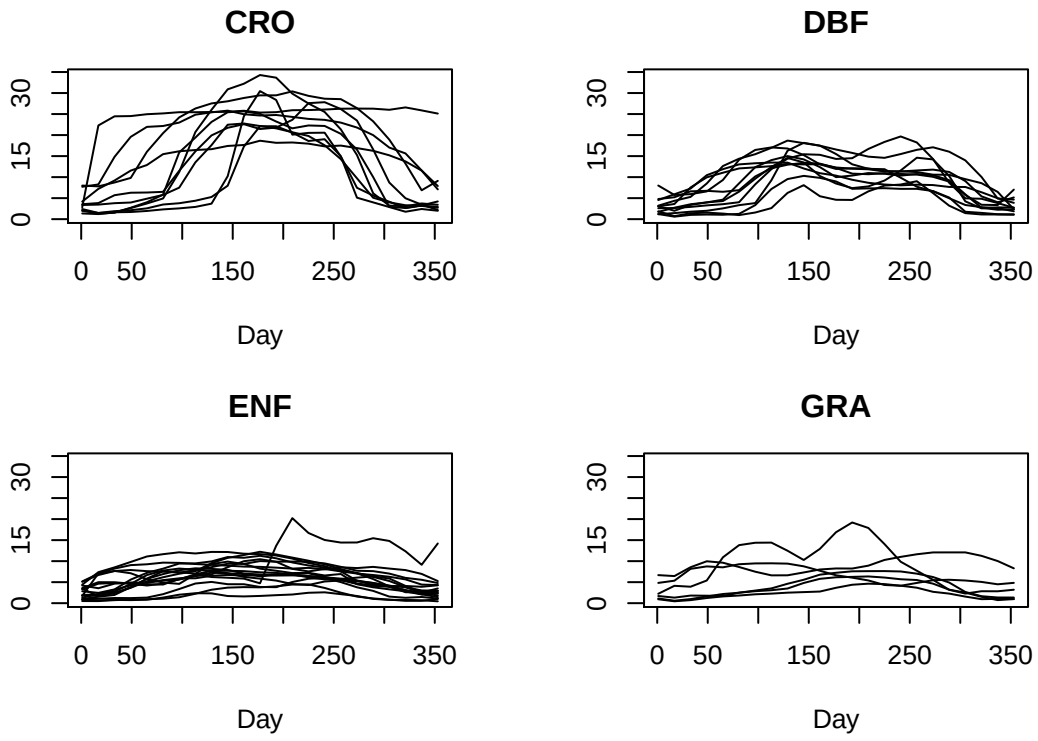


Figure 4.18: The widths of prediction interval plots of four vegetation types. For vegetation types CRO, DBF, and ENF, they have plenty of data to show that the widths of prediction intervals have the similar pattern. The curve that sharply increases after day 200 in ENF belongs to site US-Me2, whose global prediction interval can be found in Figure 4.16. As for GRA, we have not observed an obvious trend. This may due to the amount of data is not sufficient.

4.6 Discussion

We finished all the steps in Table 4.1. You may wonder why we go to the registered space, rather than directly simulate in the real time space as shown in Table 4.4. This table summarize our model steps, where y_{ij} is a discrete observation. We first smooth each year to produce $\hat{f}_i(t)$. Then we identify landmarks and apply landmark registration. We produce $\hat{f}_i^*(t)$ by smoothing points in the registered time scale. We use the fitted models of coefficients in the registered time scale to simulate new curves in the registered time scale $f^*_{simu,i}(t^*)$. Next, we use Dirichlet simulation to create new landmarks and map our curves back to the real time scale. In the end, we add noise by using the model of measurement error. The first model that we have problems took a short cut as we marked in the red arrow. The model we propose takes a detour strange using registration, because registration improves the fit.

Table 4.4: Reduced model process.

Real Time Discrete		Real Time Continuous		Registered Time Continuous
y_{ij}	$\xrightarrow[\text{Section 2.2.3}]{\text{smoothing}}$	$\hat{f}_i(t)$	$\xrightarrow[\text{Sections 2.2.3 \& 4.1}]{\text{landmark registration}}$	$\hat{f}_i^*(t^*)$
		$\downarrow$		Section 4.2 $\downarrow$ simulation
$\tilde{y}_{ij}$	$\xleftarrow[\text{Section 4.4}]{\text{add } \epsilon_{ij}}$	$f_{simu,i}(t)$	$\xleftarrow[\text{Section 4.3}]{\text{complete } h_i^{-1}(t)}$	$f^*_{simu,i}(t^*)$

To demonstrate the necessity of registration, we simulated 1000 curves using our single location model, and simulated 1000 curves without registration, only through the original coefficients in Figure 4.19. The original six-year old aspen data are plotted in color in both figures. We plot a vertical dashed line in both plots at day 150. The simulations without registration inflate the vertical variation, and the seasonal variation has not been fully captured.

The two advantages of our single location model are 1. using registration to diminish vertical variation caused by phase difference; 2. using Dirichlet regression to model landmarks, and find the range of phase variation.

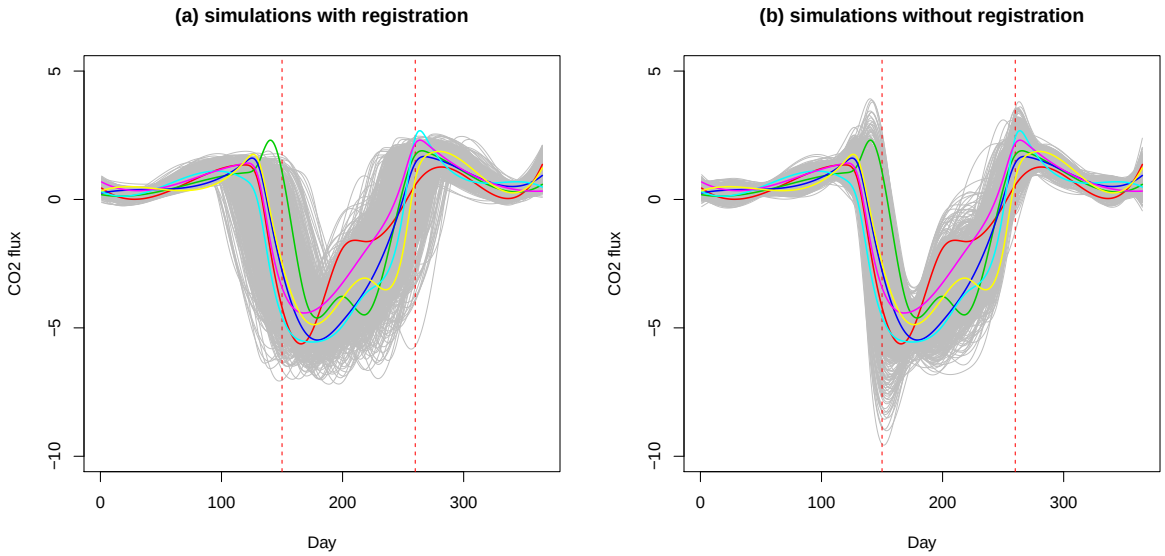


Figure 4.19: Comparison of simulations with and without registration. The smoothed observations are plotted in bright color. The 1000 simulations from different methods are plotted as grey lines in both images. Compared to the simulations on the right, the simulations with registration on the left eliminate spikes as we indicated using red vertical dashed lines.

Chapter 5

Spatial Modeling

We introduced single location modeling of CO_2 flux data in the previous chapters. In this chapter, we study multiple sites together to allow predictions at unobserved sites. We propose three methods for using nearby sites' information to build a spatial model which can predict CO_2 flux data at any location.

The first model described in Section 5.1 uses only CO_2 flux data because areas near a site may share similar variations in temperature, precipitation, illumination, etc. The forest in the same neighborhood may have similar variation in its annual CO_2 flux cycle. Because of this, the observed sites can give information about nearby unobserved ones. To achieve the goal, we study how correlation changes with distance. The results show that using only CO_2 flux data may not be sufficient. We then consider a regression type model using NDVI as the covariate. Section 5.2 introduces the functional linear regression model, where both independent and dependent variables are functional objects. Later, in Section 5.3, we consider using more information than remote sensing data, such as coordinates, elevation, and vegetation types to build a model. GAM is our option. In the end, we have a discussion of all three models in Section 5.4.

We introduce each model in the same flow: data, model, basis, results, and summary. Because of different model structures, the data used in each model may not be identical. The

data preprocessing sections present the locations of sites selected, summarizing the information of data used in each study. Next, we introduce each model. Every model uses smoothing techniques, but the bases applied are different. We use a short paragraph to describe the basis. Then we present the results. At the end of each model, we have a brief summary.

We use $y_{n_pp}(t)$ to denote the test CO_2 flux data of site p at time t. The estimation of $y_{n_pp}(t)$ is $\hat{y}_{n_pp}(t)$. The mean curve of all tests data $y_{n_pp}(t)$ is denoted by $\bar{y}_{n_pp}(t)$. For model assessment, we chose the mean square error (MSE). Besides the MSE, we use MSE(t) to track the change of MSE over time.

$$\text{MSE} = \frac{1}{\mathcal{P}}\frac{1}{T}\sum_p \sum_t \left(y_{n_pp}(t) - \hat{y}_{n_pp}(t)\right)^2, \tag{5.1}$$

$$\text{MSE}(t) \coloneqq \frac{1}{\mathcal{P}}\sum_p \left(y_{n_pp}(t) - \hat{y}_{n_pp}(t)\right)^2, \tag{5.2}$$

where $p = 1, \ldots, \mathcal{P}$, and $t = 1, \ldots, T$.

5.1 Conditional Expectation Model (CEM)

We collected data from more than 150 sites. Not all of them meet the modeling requirements. Section 5.1.1 details the data preparation process. In Section 5.1.2 we introduce our first spatial model. Inspired by single location modeling, we decompose the data into different effects/components and study each component's spatial correlation. Under certain assumptions, we propose the conditional expectation model. Section 5.1.4 displays the results including three parts: the plots of the overall fitting of each component, the predictions, and the spatial correlations.

5.1.1 Data Preparation of CEM

We use AmeriFlux sites to conduct the analysis. The number of sites we collected is 157. The information we collect includes CO_2 flux, year, day of the year, elevation, and vegetation

type. This spatial model is an extension of the single location modeling (Chapter 3 and Chapter 4), so we restrict attention to sites with complete annual CO_2 flux data. We also eliminated data in the following three situations:

- Sites from the southern hemisphere. Among 157 sites, only two were located in the southern hemisphere, where the seasons are opposite to the northern hemisphere.
- Sites south of 30°N. After plotting the data, we noticed data collected in the tropics either didn't present seasonal variation nor the seasonal pattern was very different from more northern sites.
- Sites in Alaska. There were 14 sites located in Alaska, all of which were at least 2862 *km* away from the more southern sites.

For sites located with latitude from 30°N to 57°N, some had extreme data or more outliers for other reasons. We plot the CO_2 flux of these sites in Figure 5.1. In Figure 5.1, site US-CRT has vegetation type croplands (CRO). One year out of three presented a different phase. After checking, we noticed that this site used to plant winter wheat and soybean. Different crops may cause the CO_2 flux phase differences. In our study, we removed the data from the year with the different phase. Site US-Tw3 had "wiggly" data. This site was an alfalfa field. The site description mentioned that this field is "harvested by mowing and bailing several times per year," which explains the pattern. We removed this site from the spatial study. Site US-LPH has two complete years of data. During the summertime, the measuring device was broken down around day 220. We removed this site from our study. For site US-Ha1, US-Ton, and US-Me2, one or two years of data with machine defects dragged some points far away from their underlying pattern. In our study, we removed these years of data.

After the previous filtering, the overall data time range is from 1992 to 2015. From 2000 on, more than ten sites are available each year. Our study concentrates on this subset of the data.

Our spatial study includes 89 sites, whose locations are shown in Figure 5.2. More information of these sites are listed in Table A.3. The number of sites of each vegetation type is

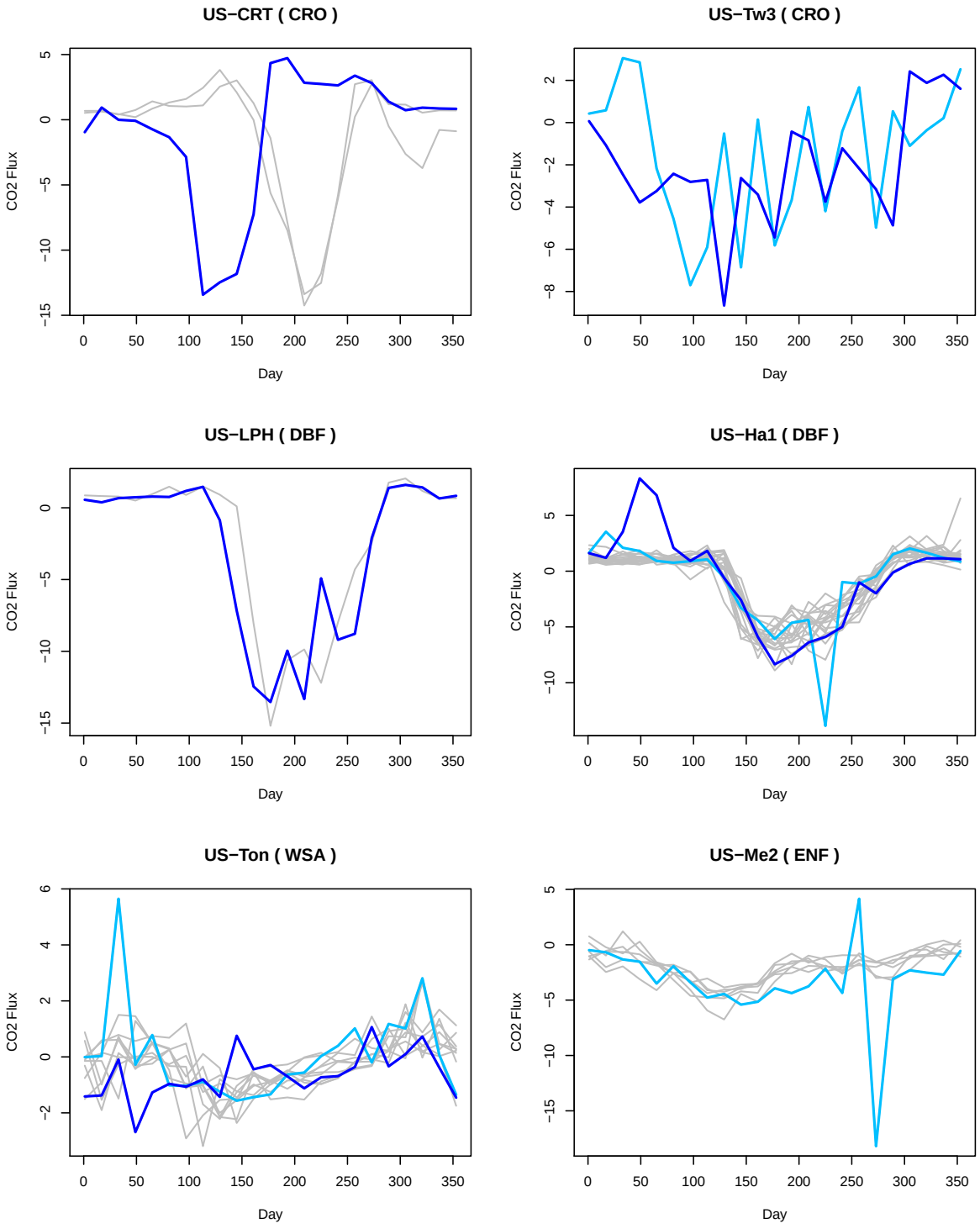


Figure 5.1: Sites with suspicious data. After checking all sites, we found these six sites may have suspicious data. We marked the suspicious data in dark or light blue.

Figure 5.2: Locations of sites used in CEM. (The squared area has latitude from 30°N to 57°N and longitude from 68.7°W to 123.6°W.)

summarized in Table 5.1. The vegetation type we have the most is ENF, followed by GRA, CRO, and DBF. The other vegetation types have no more than five sites. The time coverage of each site is shown in Figure 5.3 (a). The time range is from 2000 to 2015. Due to different research needs, the year coverages vary from sites to site. Thus, available sites change from year to year. The number of possible sites each year is shown in Figure 5.3 (b).

5.1.2 Modeling Details of CEM

We have $\mathcal{P} = 89$ sites in the system. Each site p has n_p years of data, ranging from 2 to 15. Each year has $m = 23$ observations. We use y_{ijp} to denote the j^{th} observation of year i at site p. The motivation for our spatial model is the single site model described in Chapter 3 and Chapter 4 from page 31. f denotes the underlying overall smooth effect across all sites. The smooth effect of site p is presented using ψ_p. The year effect of site p at year i is η_{ip}. The j^{th}

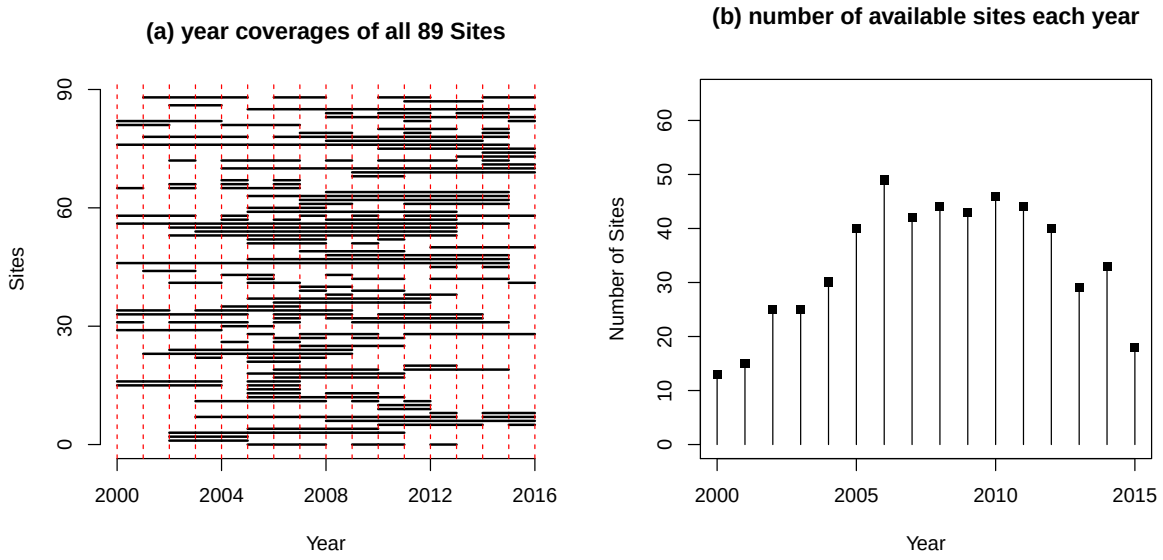


Figure 5.3: Year coverage of sites and number of sites each year. Image (a) plots the year coverage for each site. Image (b) summarize the number of sites in each year.

Table 5.1: Summary of number of sites for each species.

	Vegetation Type	**Number**
CRO	Croplands	13
CSH	Closed shrublands	4
DBF	Deciduous broadleaf forests	12
ENF	Evergreen needleleaf forest	26
GRA	Grasslands	17
MF	Mixed forest	5
OSH	Open shrublands	4
SAV	Savannas	1
WET	Permanent wetlands	4
WSA	Woody savannas	3
Total		89

measurement error of year i at site p is ϵ_{ijp}. This gives the spatial model:

$$f_{ip} = f + \psi_p + \eta_{ip}, \tag{5.3}$$

$$= f_p + \eta_{ip}, \tag{5.4}$$

$$y_{ijp} = f(t_j) + \psi_p(t_j) + \eta_{ip}(t_j) + \epsilon_{ijp}, \tag{5.5}$$

$$= f_p(t_j) + \eta_{ip}(t_j) + \epsilon_{ijp}, \tag{5.6}$$

where $i = 1, \ldots, n_{\mathcal{P}}$, $j = 1, \ldots, m$, and $p = 1, \ldots, \mathcal{P}$.

Model (5.5) differs from (3.9) in the distributional assumptions in the following ways.

In single location modeling, for a fixed location, we assumed the measurement errors ϵ_{ij*} were independent.

$$\epsilon_{(ijp|p=*)} \coloneqq \epsilon_{ij*} \overset{\text{independent}}{\sim} N(0, \sigma^2_{\epsilon_{ij*}}) \tag{5.8}$$

In multi-location modeling, sites affect each other. Our assumption of ϵ_{ijp} is

$$\epsilon_{ijp} \overset{\text{identical}}{\sim} N(0, \sigma^2), \tag{5.9}$$

and the joint distribution of $(\epsilon_{ij1}, \epsilon_{ij2}, \ldots, \epsilon_{ijp})$ is assumed to be

$$\begin{pmatrix} \epsilon_{ij1} \\ \epsilon_{ij2} \\ \vdots \\ \epsilon_{ij\mathcal{P}} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{\mathcal{P}} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1\mathcal{P}}\sigma_1\sigma_{\mathcal{P}} \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \ldots & \rho_{2\mathcal{P}}\sigma_2\sigma_{\mathcal{P}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1\mathcal{P}}\sigma_1\sigma_{\mathcal{P}} & \rho_{2\mathcal{P}}\sigma_2\sigma_{\mathcal{P}} & \ldots & \sigma_{\mathcal{P}}^2 \end{pmatrix}\right) \tag{5.10a}$$

$$= MVN\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho_{12}\sigma^2 & \ldots & \rho_{1\mathcal{P}}\sigma^2 \\ \rho_{12}\sigma^2 & \sigma^2 & \ldots & \rho_{2\mathcal{P}}\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1\mathcal{P}}\sigma^2 & \rho_{2\mathcal{P}}\sigma^2 & \ldots & \sigma^2 \end{pmatrix}\right). \tag{5.10b}$$

In the following context, we use $\mathcal{S}$ to denote a certain smoother, and the coefficients $\boldsymbol{\beta}$ of smooth processes are all calculated using least squares. We model each term of (5.5) in the following order:

1. We first model the smooth effect of each year i at each location p by fitting a least squares

model for all the m observations of the i^{th} year at location p.

$$f_{ip} = \mathcal{S}(y_{ijp} : \forall j = 1, \ldots, m \mid i, p) \tag{5.11}$$

$$= \Phi_{f_{ip}} \boldsymbol{\beta}_{f_{ip}} \tag{5.12}$$

2. According to (5.7), the measurement error is calculated by

$$\epsilon_{ijp} = y_{ijp} - f_{ip}(t_j). \tag{5.13}$$

3. Next, for each site p, we estimate the underlying smooth process for location p by averaging all years of smooth effects at site p.

$$f_p = \frac{1}{n_p} \sum_{i=1}^{n_p} f_{ip} \tag{5.14}$$

$$= \Phi_{f_p} \boldsymbol{\beta}_{f_p} \tag{5.15}$$

4. Based on (5.6) and (5.7), we estimate the year effect η_{ip} of the i^{th} year at site p.

$$\eta_{ip} = f_{ip} - f_p \tag{5.16}$$

$$= \Phi_{f_{ip}} \boldsymbol{\beta}_{f_{ip}} - \Phi_{f_p} \boldsymbol{\beta}_{f_p} \tag{5.17}$$

$$= \Phi_{\eta_{ip}} \boldsymbol{\beta}_{\eta_{ip}} \tag{5.18}$$

5. The overall underlying effect f across all sites is estimated by averaging the underlying smooth effect of each site.

$$f = \frac{1}{\mathcal{P}} \sum_{p=1}^{\mathcal{P}} f_p, \tag{5.19}$$

$$= \Phi_f \boldsymbol{\beta}_f \tag{5.20}$$

6. The site effect ψ_p is calculated by

$$\psi_p = f_p - f \tag{5.21}$$

$$= \Phi_{f_p}\boldsymbol{\beta}_{f_p} - \Phi_f\boldsymbol{\beta}_f \tag{5.22}$$

$$= \Phi_{\psi_p}\boldsymbol{\beta}_{\psi_p}. \tag{5.23}$$

For each fixed location, we estimated the above six terms and recorded three landmarks for each annual data f_{ip}.

In Figure 5.4 we demonstrate how we identify three landmarks. We marked two maximal values at spring and fall, and one minimal value during summer.

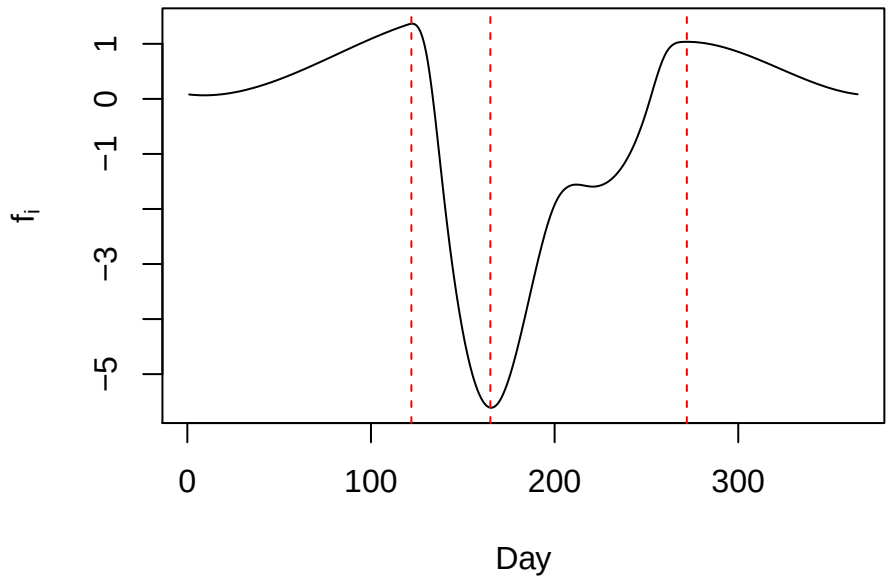


Figure 5.4: Identifying landmarks on hypothetical f_{ip} curve. The vertical red dashed lines indicate the location of landmarks - two maximal and one minimal.

Spatial Correlation

For ϵ_{ijp}, $\eta_{ip}(t_j)$, $\psi_p(t_j)$ and landmarks, we are interested in their spatial correlations.

To calculate the linear correlation of two sites, we consider the data from the same time period. The time information available varies for different variables as listed in Table 5.2. There are three types of time availability:

1. Data have both year and day information, such as ϵ_{ijp} and $\eta_{ip}(t_j)$. The collection of available time of site p is denoted by

$$T_{p(Year,Day)} = \{(Year, Day) : Year = 1; \ldots, n_p,\ Day = 1; \ldots, 23,\ \text{Site} = p\}. \tag{5.24}$$

2. Data have only year information, for example, annual landmarks. The collection of available time of site p is

$$T_{p(Year)} = \{(Year) : Year = 1, \ldots, n_p;\ \text{Site} = p\}. \tag{5.25}$$

3. Data have only day information, like $\psi_p(t_j)$. Its collection of available time is

$$T_{p(Day)} = \{(Day) : Day = 1, \ldots, 23;\ \text{Site} = p\}. \tag{5.26}$$

Table 5.2: Time availability of ϵ_{ijp},η_{ip},ψ_p, and landmark.

	Year	**Day**	**Time Collection**	**Length per Location per Year**
ϵ_{ijp}	✓	✓	$T_{p(Year,Day)}$	23
$\eta_{ip}(t_j)$	✓	✓	$T_{p(Year,Day)}$	23
$\psi_p(t_j)$	✗	✓	$T_{p(Day)}$	23
Annual Landmark	✓	✗	$T_{p(Year)}$	3

We assume site p_1 has time collection T_{p_1} and site p_2 has time collection T_{p_2}, i.e., observations are available in these two periods. Their overlapped time region $T = T_{p_1} \cap T_{p_2}$ belongs to one of the three categories: $T_{(Year,Day)}$, $T_{(Year)}$, and $T_{(Day)}$. With our data in hand, T will consist of zero or more whole years, but in other studies or with different selection criteria, other overlaps would be possible.

We use $\boldsymbol{x}$ to denote a variable of interest, one of ϵ_{ijp}, $\eta_{ip}(t_j)$, $\psi_p(t_j)$ and the landmarks. Let x_{ijp} represent the data of the i^{th} year the j^{th} observation at site p.

We define the collection of pairs of non-identical sites as

$$\boldsymbol{P}_{pairs} = \{(p_1, p_2) : \forall p_1, p_2 \in [1, \ldots, \mathcal{P}] \text{ and } p_1 \neq p_2\} \tag{5.27}$$

For each pair of sites $(p_1, p_2) \in \boldsymbol{P}_{pairs}$, at their overlapped time region T, we assume the linear correlation is a function of the distance:

$$d = distance(p_1, p_2), \tag{5.28}$$

$$\rho(d) = cor(\boldsymbol{x}_{\boldsymbol{p_1}}, \boldsymbol{x}_{\boldsymbol{p_2}}), \tag{5.29}$$

where both $\boldsymbol{x}_{\boldsymbol{p_1}}$ and $\boldsymbol{x}_{\boldsymbol{p_2}}$ are restricted to the overlapped time region T.

To estimate the relation between d and $\rho(d)$, we calculated the observed value of $(\boldsymbol{\rho}, \boldsymbol{d})$ for all pairs of sites, and fitted them using an exponential decay function

$$\rho = ce^{-\lambda d}, \text{ where } c, \lambda > 0, \tag{5.30}$$

for ϵ_{ijp}, $\eta_{ip}(t_j)$, $\psi_p(t_j)$ and landmarks separately.

Conditional Expectation Modeling

Let site 1 be our site of interest. We assume the overall smooth effect across all sites f, all the site effects ψ_p, and all the year effects η_{ip} are fixed. The only random effect of the multilocation modeling is the measurement error ϵ_{ijp}.

Single Observation Update

To shorten the notation, we define

$$y_{ij\boldsymbol{p}^*} \coloneqq (y_{ij2}, y_{ij3}, \ldots, y_{ij\mathcal{P}}), \tag{5.31}$$

$$\epsilon_{ij\boldsymbol{p}^*} \coloneqq (\epsilon_{ij2}, \epsilon_{ij3}, \ldots, \epsilon_{ij\mathcal{P}}). \tag{5.32}$$

The annual average CO_2 flux data of site 1 at year i day j given the other $(\mathcal{P}-1)$ sites is

$$\boldsymbol{E}(y_{ij1}|y_{ij\boldsymbol{p}^*}) = \boldsymbol{E}(y_{ij1}|y_{ij2}, y_{ij3}, \ldots, y_{ij\mathcal{P}}), \tag{5.33}$$

$$= f(t_j) + \psi_1(t_j) + \eta_{i1}(t_j) + \boldsymbol{E}(\epsilon_{ij1}|\epsilon_{ij2}, \epsilon_{ij3}, \ldots, \epsilon_{ij\mathcal{P}}), \tag{5.34}$$

$$= f(t_j) + \psi_1(t_j) + \eta_{i1}(t_j) + \boldsymbol{E}(\epsilon_{ij1}|\epsilon_{ij\boldsymbol{p}^*}). \tag{5.35}$$

To help calculate $\boldsymbol{E}(\epsilon_{ij1}|\epsilon_{ij\boldsymbol{p}^*})$, we define two matrices. The first one is the $(\mathcal{P}-1)\times(\mathcal{P}-1)$ variance-covariance matrix of $\epsilon_{ij\boldsymbol{p}^*}$ denoted as $\Sigma_{\boldsymbol{p}^*}$, the other is a $1\times(\mathcal{P}-1)$ matrix Σ_B whose elements are the covariances between ϵ_{ij1} and each element of $\epsilon_{ij\boldsymbol{p}^*}$.

$$\boldsymbol{E}(\epsilon_{ij1}|\epsilon_{ij\boldsymbol{p}^*}) = \mu_1 + \Sigma_B\Sigma_{\boldsymbol{p}^*}^{-1}\begin{bmatrix}\epsilon_{ij2}-\mu_2\\ \epsilon_{ij3}-\mu_3\\ \vdots\\ \epsilon_{ij\mathcal{P}}-\mu_{\mathcal{P}}\end{bmatrix}, \tag{5.36a}$$

$$= \Sigma_B\Sigma_{\boldsymbol{p}^*}^{-1}\begin{bmatrix}\epsilon_{ij2}\\ \epsilon_{ij3}\\ \vdots\\ \epsilon_{ij\mathcal{P}}\end{bmatrix}. \tag{5.36b}$$

The expectation of $\boldsymbol{E}(\epsilon_{ij1}|\epsilon_{ij\boldsymbol{p}^*})$ is 0.

Annual Observations Update

We use $y_{i.1}$ and $\epsilon_{i.1}$ to denote annual data and errors at location 1 year i.

$$y_{i.1} \coloneqq (y_{i11}, y_{i21}, \ldots, y_{im1}) \tag{5.37}$$

$$\epsilon_{i.1} \coloneqq (\epsilon_{i11}, \epsilon_{i21}, \ldots, \epsilon_{im1}) \tag{5.38}$$

The collection of annual data at other locations is denoted as

$$\epsilon_{i.\boldsymbol{p}^*} \coloneqq (\epsilon_{i1\boldsymbol{p}^*}, \epsilon_{i2\boldsymbol{p}^*}, \ldots, \epsilon_{im\boldsymbol{p}^*}). \tag{5.39}$$

The conditional expectation of the annual errors is

$$\boldsymbol{E}(\epsilon_{i.1}|\epsilon_{i.\boldsymbol{p}^*}) = \Sigma_B \Sigma_{\boldsymbol{p}^*}^{-1} \begin{bmatrix} \epsilon_{i12} & \epsilon_{i22} & \ldots & \epsilon_{im2} \\ \epsilon_{i13} & \epsilon_{i23} & \ldots & \epsilon_{im3} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{i1\mathcal{P}} & \epsilon_{i2\mathcal{P}} & \ldots & \epsilon_{im\mathcal{P}} \end{bmatrix}, \tag{5.40}$$

whose expectation is $\mathbf{0}$.

Underlying Annual Effect Update

According to (5.35), the modified underlying smooth annual effect at site 1 year i is:

$$\tilde{f}_{i1} = \mathcal{S}(y_{i.1} - \boldsymbol{E}(\epsilon_{i.1}|\epsilon_{i.\boldsymbol{p}^*})), \tag{5.41}$$

where $\mathcal{S}$ is a smoother.

Because the smoother $\mathcal{S}$ we chose uses OLS, which is linear, (5.41) can be written as:

$$\tilde{f}_{i1} = \mathcal{S}(y_{i.1}) - \mathcal{S}(\boldsymbol{E}(\epsilon_{i.1}|\epsilon_{i.\boldsymbol{p}^*})). \tag{5.42}$$

Though $\boldsymbol{E}(\epsilon_{i.1}|\epsilon_{i.\boldsymbol{p}^*})$ helps to update observations; unfortunately, the updating of the annual effect, $\mathcal{S}(\boldsymbol{E}(\epsilon_{i.1}|\epsilon_{i.\boldsymbol{p}^*}))$ is approximately 0 at each time point.

Unobserved Site Prediction

For unobserved sites, we cannot directly use the CEM. However, we still want to see the "out of sample" predictions. We cheated a bit. Here is what we did. As our model (5.5) describes, the flux data is composed of the underlying overall smooth effect f, the site effect ψ, the year effect η, and measurement error ϵ. The f can be estimated by the training dataset as described in (5.19). For the other three terms, we use the same concept when calculating the conditional expectation of measurement error (5.36b). When we calculate the variance-covariance matrix of each component, we used the corresponding testing data.

5.1.3 Basis Used in CEM

Different from the previous chapters, the basis we used in this chapter in the smoothing process is the cyclic cubic spline, where values at the start and end of the year match up to the second derivative. The reasons we chose this basis are:

- data used in the spatial model has the property that data has low variation and tends to zero during the beginning and the end of a year,
- using the cyclic basis, the regression fit for the starting and ending points has less variation. The smoothed data used in the spatial analysis using the non-cyclic basis and the cyclic basis are displayed in Figure 5.5.

This simplification would not work in the tropics and the southern hemisphere.

5.1.4 CEM Results

Fitting Results Overview

We use $\hat{f}$, $\hat{\psi}_p$, $\hat{\eta}_{ip}$ and e_{ijp} to denote the estimates of f, ψ_p, η_{ip}, and ϵ_{ijp}. The basis of all the fitting is the same cyclic basis with knots (0,120,130,140,170,200,220,250,260,270,365).

To compare the amplitudes of $\hat{f}$, e_{ijp}, $\hat{\psi}_p$ and $\hat{\eta}_{ip}$, we plot them in Figure 5.6.

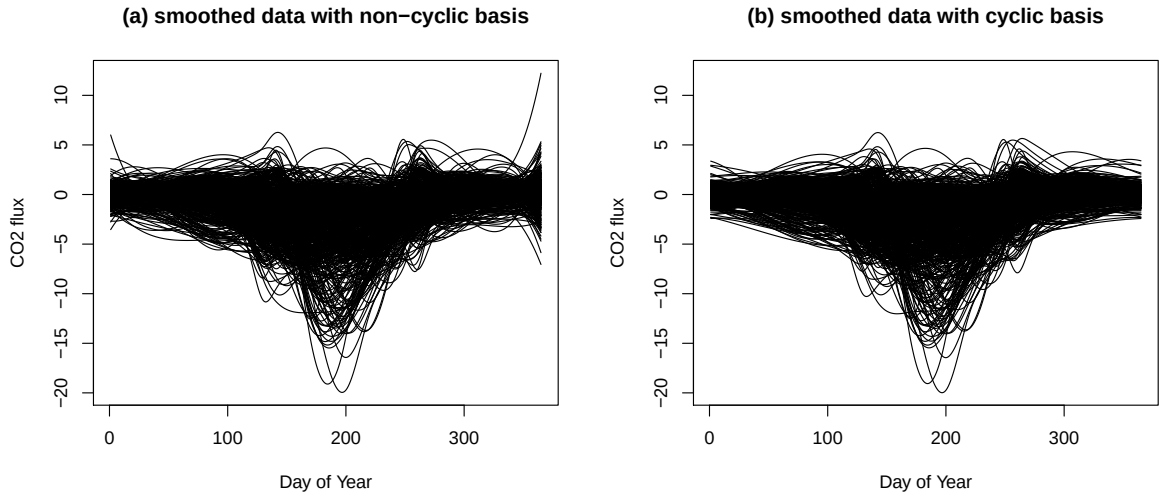


Figure 5.5: Non-cyclic and cyclic basis comparison. Image (a) plots the smoothed curves using non-cyclic basis. Image (b) presents the smoothed curves using cyclic basis. The fitted curves in both images are very similar in the middle of the year, but quite different near the ends. The ends of the smoothed curves using cyclic basis are less spread.

Predictions

The parameter we used to estimate the $\rho(d)$ is summarized in Table 5.3. The details of the parameter estimations are shown at the end of this section.

Table 5.3: Parameters used in $\rho(d)$.

Component	c	λ
ψ	0.476	0.00204
η	0.171	0.00165
ϵ	0.318	0.00250

We plot partial predictions of CEM in Figure 5.7. Since every plot has the same y limit, it's obvious to see there is not much variation in the predictions.

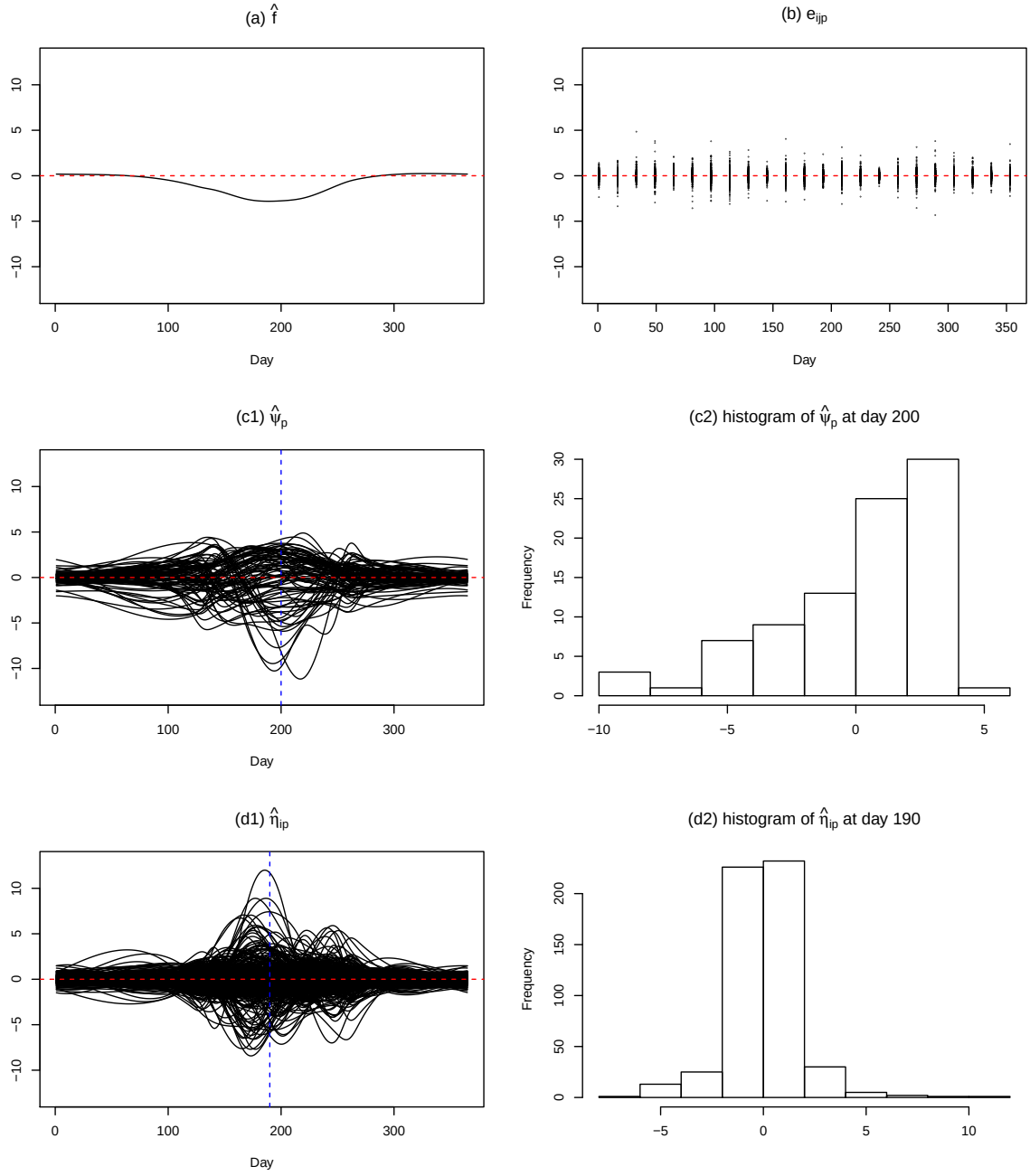


Figure 5.6: Spatial model fitting results. Image (a) plots $\hat{f}$, the estimation of the overall smooth effect across all sites. From day 70 to day 280, this effect increases from zero to the top around day 200 and then diminishes back to zero. During the whole wintertime, this effect is approximately zero. The maximum amplitude of $\hat{f}$ is around three. Image (b) displays the residuals e_{ijp}. They distribute along the x-axis. Image (c1) shows the estimations of 89 site effects $\hat{\psi}_p$, whose maximal amplitude is around 15. Data with the most variation are around day 200. Image (c2) shows the histogram of $\psi_p(t_j), p = 1, \ldots, 89, t_j = 200$. The distribution is skewed to the left. Image (d1) plots the 536 year effects η_{ip}. The maximal amplitude is up to 20. The most variation is around day 190. Image (d2) is the histogram of 536 $\eta_{ip}(t_j)$ at day 190. This histogram is more symmetric than (c2).

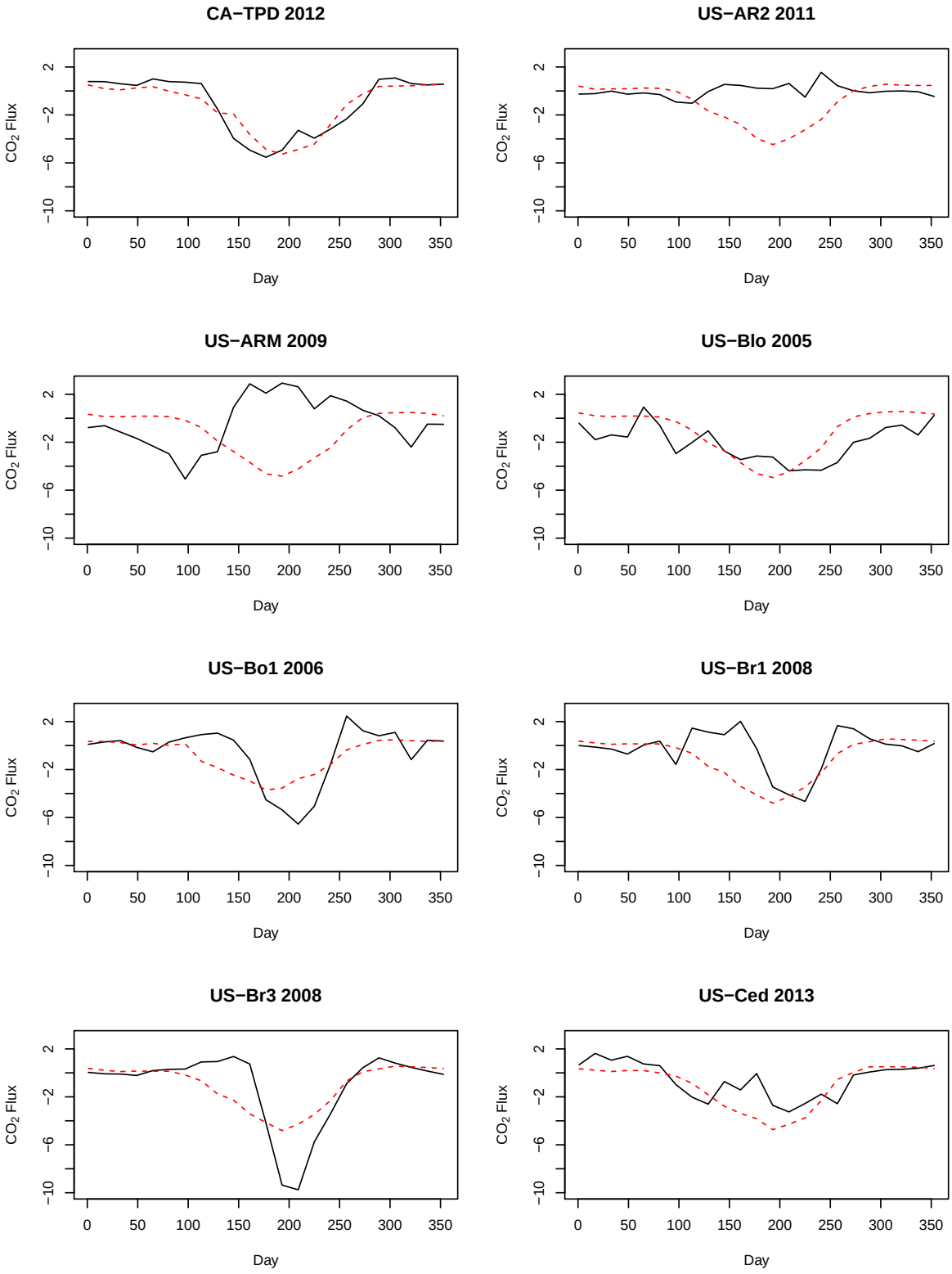


Figure 5.7: Predictions from CEM. The black solid lines are observations, the red solid lines are predictions. We see there are variations in the observations; however, the predictions are quite the same.

Correlation of Each Component

Even though CEM was not very successful, it is still worth looking at the spatial correlation of each component.

1. **Residuals** e_{ijp}

   For every location, each year, we use spline smoothing to estimate f_{ip}. We calculate the residual e_{ijp} using:

   $$e_{ijp} = y_{ijp} - \hat{f}_{ip}(t_j). \tag{5.43}$$

   The plots of the residuals are shown in Figure 5.8.

   The residuals of each site at the overlapped time region are

   $$\boldsymbol{x}_{p_1} \coloneqq \left\{ e_{ijp} : (i,j) \in T_{(Year,\, Day)}, p = p_1 \right\}, \tag{5.44}$$

   $$\boldsymbol{x}_{p_2} \coloneqq \left\{ e_{ijp} : (i,j) \in T_{(Year,\, Day)}, p = p_2 \right\}. \tag{5.45}$$

   The linear correlation is calculated as

   $$\boldsymbol{\rho}_{e_{ijp}} = \left\{ \rho : cor(\boldsymbol{x}_{p_1}, \boldsymbol{x}_{p_2}), \forall (p_1, p_2) \in \boldsymbol{P}_{pairs} \right\}. \tag{5.46}$$

   The parameters in (5.30) were estimated from these data using non-linear least squares in R. The function $\rho(d)$ was also estimated non-parametrically using R function `npreg` in package `np` (Hayfield et al., 2008).

   The results of the estimation of (5.30) are shown in Table 5.4 and Figure 5.9.

Table 5.4: Nonlinear least squares estimation of ϵ_{ijp}.

	Estimate	Std. Error
c	0.318	0.027
λ	0.00250	0.00027

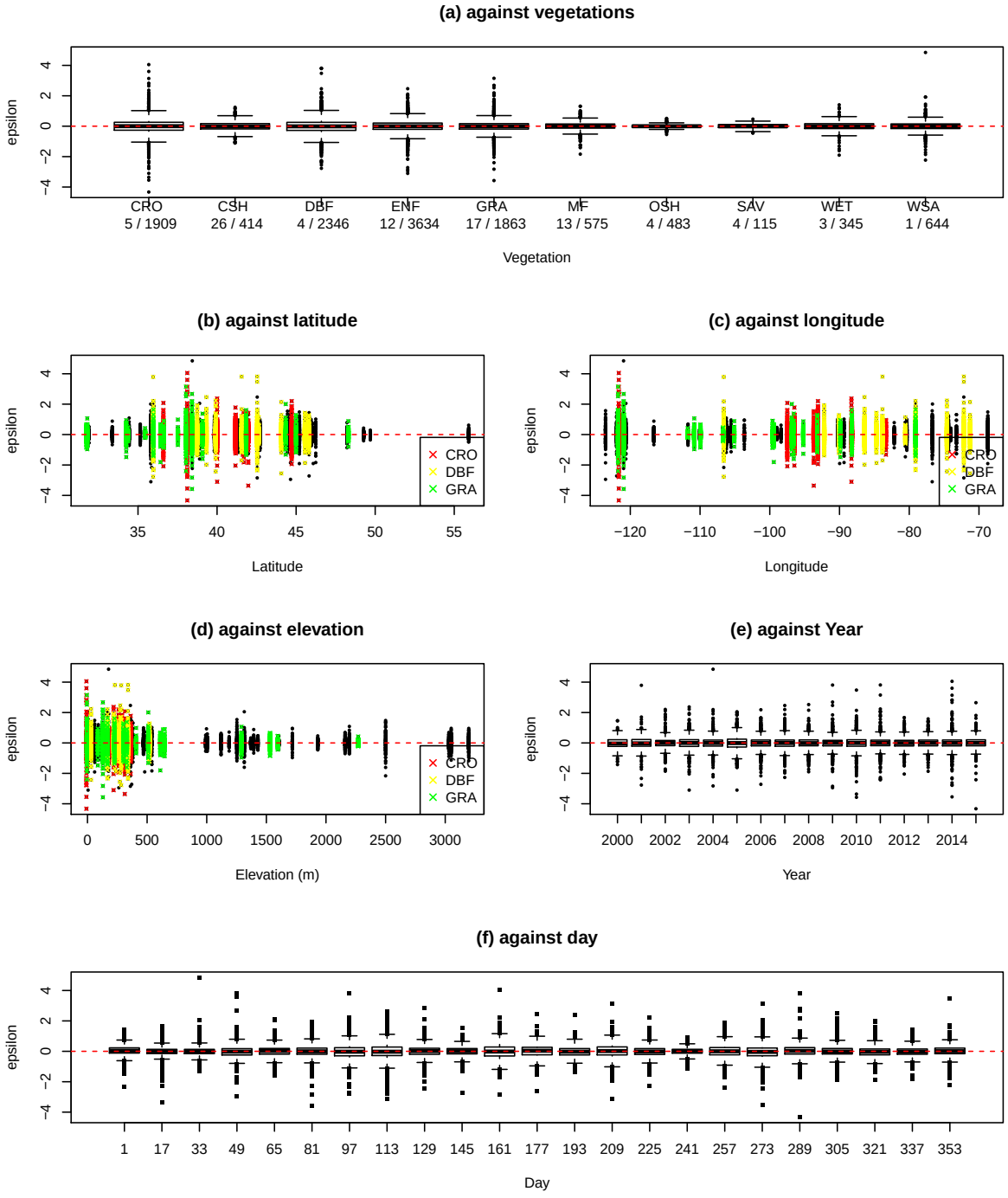


Figure 5.8: ϵ_{ijp} against each variable. Image (a) displays the residual against vegetation. The medians of residuals of each vegetation are around zero. On the x-axis we have the name of each vegetation together with the number of sites and the number of points that form each box plot. Some vegetation types have more data than others, for example, ENF has 12 sites and 3634 data points, while SAV has only 4 sites and 115 data points. Images (b) and (c) present the residual against latitude and longitude respectively. The majority of sites lie in an area where latitude ranges from 35°N to 50°N. (d) shows the residual against elevation in meter. Image (e) plots the residual against year. (f) displays the residual against day.

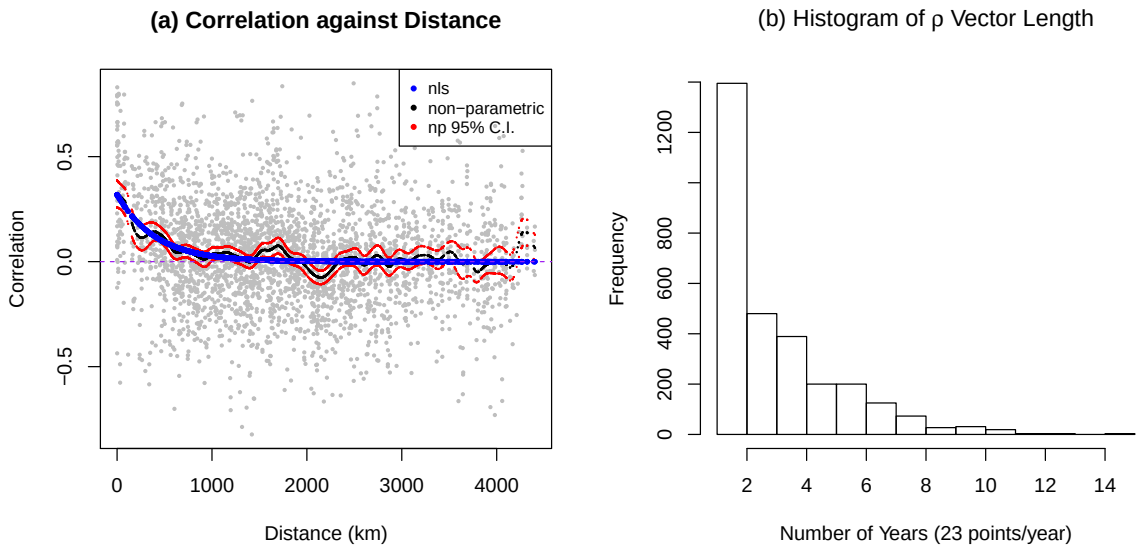


Figure 5.9: How $\epsilon'_{ijp}s$ correlation changes with distance and the length of the vectors used to calculate the correlations. Image (a) displays the correlation against distance. The blue dots are the fitted exponential decay line, where the correlation quickly decays to zero around 1000 *km*. The black dots are the fitting from non-parametric method and the red dots are the corresponding 95% confidence interval. Image (b) shows the histogram of the length of the vector used to calculate the correlation in years. The majority of pairs overlap in one or two years. We have 23 residual points every year.

2. **Year Effect** η_{ip}

To calculate the year effect, we first calculate the overall smooth effect f_p of each site. The underlying overall smooth process f_p for site p is fitted using:

$$\hat{f}_p = \frac{1}{n_p} \sum_{i=1}^{n_p} \hat{f}_{ip} = \Phi_{f_p} \hat{\boldsymbol{\beta}}_{f_p}, \tag{5.47}$$

where the vector $\hat{\boldsymbol{\beta}}_{f_p}$ is calculated by least squares.

We considered two methods to fit f_p:

- fitting smooth for each year $y_i = \Phi\beta_i$ and take the average of all the fitted coefficients $Y^* = \Phi\frac{\sum\beta^*}{n}$;
- taking the average of multiple years of observations $\tilde{Y} = \frac{\sum y_i}{n}$ and fit the average observations $\tilde{Y} = \Phi\tilde{\beta}$.

Because of the linearity of our smoother, the two methods lead to the same results. We chose the latter one.

The $\eta_{ip}(t_j)$ is calculated using

$$\hat{\eta}_{ip}(t_j) = \hat{\hat{f}}_{ip}(t_j) - \hat{f}_p(t_j). \tag{5.48}$$

We plot the year effect against other variables in Figure 5.10.

The $\hat{\eta}_{ip}(t_j)$ of two sites at the overlapped time region are

$$\boldsymbol{x}_{p_1} \coloneqq \left\{\hat{\eta}_{ip}(t_j) : (i,j) \in T_{(Year,\, Day)}, p = p_1\right\}, \tag{5.49}$$

$$\boldsymbol{x}_{p_2} \coloneqq \left\{\hat{\eta}_{ip}(t_j) : (i,j) \in T_{(Year,\, Day)}, p = p_2\right\}. \tag{5.50}$$

The linear correlation is calculated as

$$\boldsymbol{\rho}_{\eta_{(t)}} = \left\{\rho : cor(\boldsymbol{x}_{p_1}, \boldsymbol{x}_{p_2}), \forall (p_1, p_2) \in \boldsymbol{P}_{pairs}\right\}. \tag{5.51}$$

We summarize the exponential decay model of $\eta(t_j)$ in Table 5.5 and Figure 5.11.

Table 5.5: Nonlinear least squares estimation of $\eta_{ip}(t_j)$.

	Estimate	Std. Error
c	0.171	0.025
λ	0.00165	0.00029

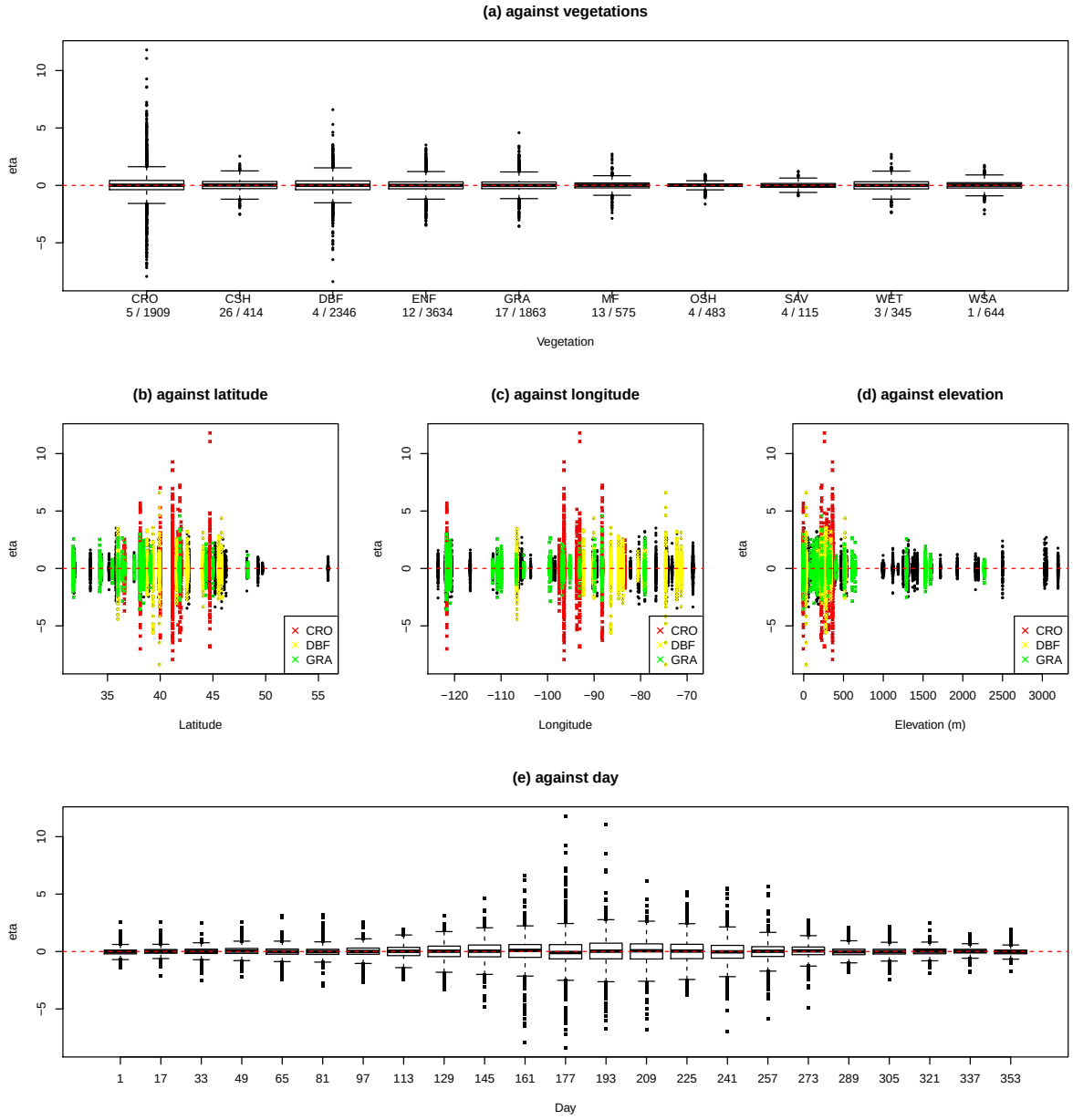


Figure 5.10: $\eta_{ip}(t_j)$ against each variable. Image (a) indicates that CRO has the most outliers. Images (b) to (d) shows CRO sites's locations. Their elevations are all below 500 m. Image (e) plots $\eta_{ip}(t_j)$ against the day of year. The medians are around zero. Data have higher variation in the summertime.

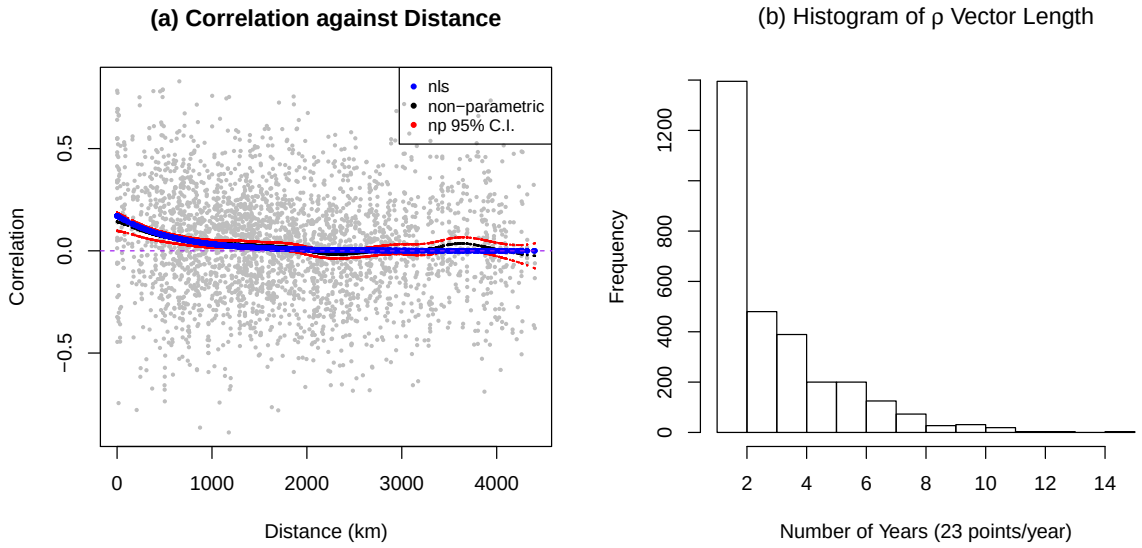


Figure 5.11: How $\eta_{ip}(t_j)'s$ correlation changes with distance and the length of the vectors used to calculate the correlations. Image (a) displays the correlation against distance. The blue dots are the fitted exponential decay line, where the correlation quickly decays to zero around 1000 *km*. The black dots are the fitting from non-parametric method and the red dots are the corresponding 95% confidence interval. Image (b) shows the histogram of the length of the vector used to calculate the correlation in years. The majority of pairs overlap in one or two years. We have 23 residual points every year.

3. **Site Effect ψ_p**

The underlying overall smooth process f across all sites is fitted using:

$$\hat{f} = \frac{1}{\mathcal{P}} \sum_{p=1}^{\mathcal{P}} \hat{f}_p = \Phi_f \hat{\boldsymbol{\beta}}_f. \tag{5.52}$$

The site effect $\psi_p(t_j)$ is estimated using:

$$\hat{\psi}_p(t_j) = \hat{f}_p(t_j) - \hat{f}(t_j). \tag{5.53}$$

We plot the estimated site effect against other variables in Figure 5.12.

The $\hat{\psi}_p(t_j)$ of two sites at the overlapped time region are

$$\boldsymbol{x}_{p_1} \coloneqq \left\{\hat{\psi}_p(t_j) : j \in T_{(Day)}, p = p_1\right\}, \tag{5.54}$$

$$\boldsymbol{x}_{p_2} \coloneqq \left\{\hat{\psi}_p(t_j) : j \in T_{(Day)}, p = p_2\right\}. \tag{5.55}$$

The spatial correlation of $\psi_p(t_j)$ is calculated as:

$$\rho_{\psi_p(t_j)} = \left\{\rho : cor(\boldsymbol{x}_{p_1}, \boldsymbol{x}_{p_2}), \forall (p_1, p_2) \in \boldsymbol{P}_{pairs}\right\}. \tag{5.56}$$

Table 5.6 and Figure 5.13 show the spatial correlation estimations.

Table 5.6: Nonlinear least squares estimation of $\psi_p(t_j)$.

	Estimate	Std. Error
c	0.476	0.055
λ	0.00204	0.00028

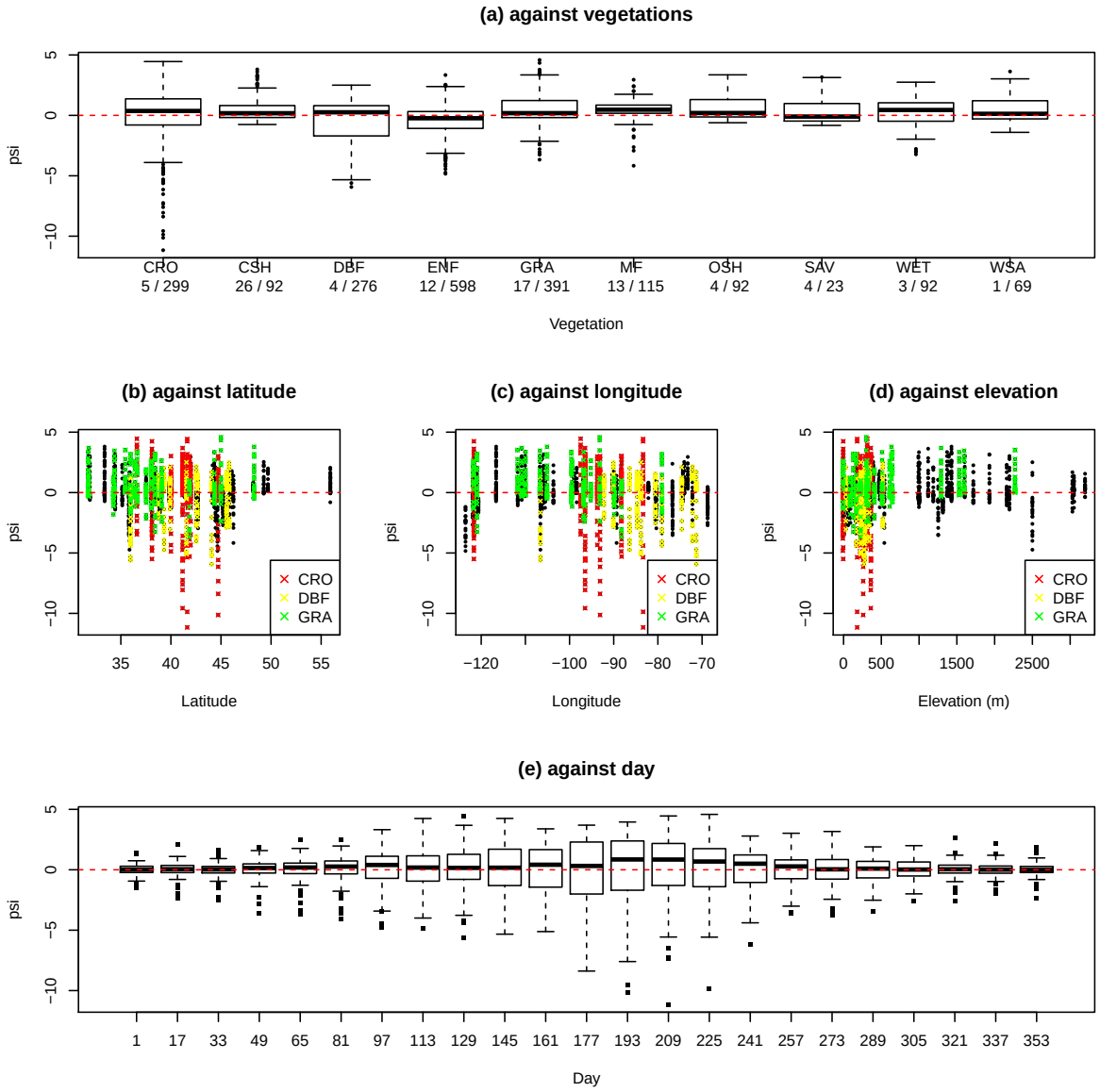


Figure 5.12: $\psi_p(t_j)$ against each variable. Image (a) shows CRO and DBF has the most variation. CRO also has the most outliers. Images (b) to (d) we plot CRO in red, DBF in yellow, and GRA in green. CRO and DBF have similar latitudes ranging from 35°N to 40°, while GRA covers a wider range of latitudes ranging from 30°N to 47°. As for longitudes, CRO and GRA cover 120°W to 80°W, and DBF ranges from 110°W to 70W. When it comes to elevation, sites with vegetation types CRO and DBF all have elevation under 1000 m. Sites with vegetation type GRA have elevations from 0 to 2300 m. Image (e) pictures the shape and variation of ψ_p. The variation of ψ_p increases from day 1 to day 193 and decrease from day 193 to the end. The medians are below 0 during summertime.

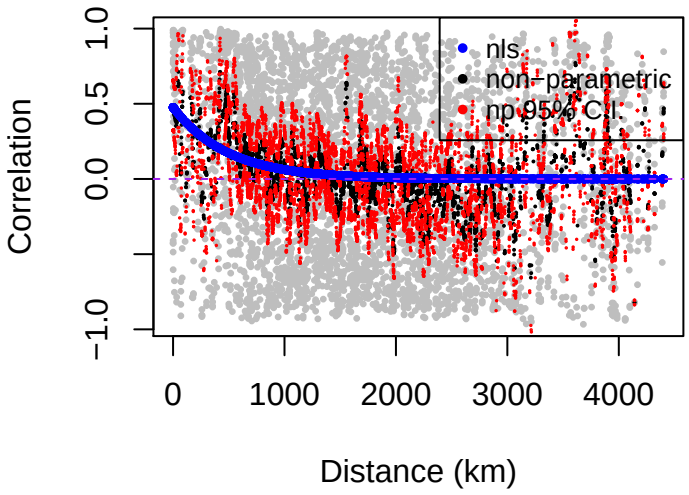


Figure 5.13: How $\psi_p(t_j)'s$ correlation changes with distance. The blue dots are the fitted exponential decay line. The black dots are the fitting from non-parametric method and the red dots are the corresponding 95% confidence interval.

4. **Landmarks**

   We are interested in how the three landmarks change across different vegetation, how the three landmarks change across multiple consecutive years and the spatial correlation of each landmark.

   We plot the three landmarks against species in Figure 5.14. Figure 5.15 displays landmarks against year.

   We use $\boldsymbol{lm}_{ipk}$ to denote the k^{th} landmarks detected from site p of year i.

   For each landmark of two sites at their overlapped years are

$$\boldsymbol{x}_{p_1} \coloneqq \{\boldsymbol{lm}_{ipk} : i \in T_{(Year)}, p = p_1\}, \tag{5.57}$$

$$\boldsymbol{x}_{p_2} \coloneqq \{\boldsymbol{lm}_{ipk} : i \in T_{(Year)}, p = p_2\}. \tag{5.58}$$

   The linear correlation is calculated as

$$\boldsymbol{\rho}_{\boldsymbol{lm}_{ipk}} = \{\rho : cor(\boldsymbol{x}_{p_1}, \boldsymbol{x}_{p_2}, \forall (p_1, p_2) \in \boldsymbol{P}_{pairs}\} \tag{5.59}$$

Figure 5.16 to Figure 5.18, their subplots (a) display the correlation against distance. The blue dots are the fitted exponential decay line. The distances of correlation decay to zero from landmark one to landmark three are around 4000 km, 2500 km, and 2000 km. Figure 5.16 to Figure 5.18, subplots (b) shows the histogram of the length of the vector used in correlation calculation. Though for each landmark we only have one dot per year, when we calculate the correlation of landmarks, we required at least three different landmarks in each landmark vector, which makes the histograms of ρ vector length are not identical for three landmarks. For example, we have two sites, which have three years of overlaps. The detected landmark vector for each landmark is shown in Table 5.7. For the first landmark, because the length of the unique landmark in both sites are less than three, so we exclude this record. For the second landmark, the two vectors meet our requirement. The correlation between the two vector is recorded. For the third landmark, the vector from site A only has two different landmarks, so

Table 5.7: Landmark vector calculation example.

Landmark	**Site A**	**Site B**	
landmark 1	(10,15,10)	(12,10,10)	✗
landmark 2	(175,160,180)	(160,170,190)	✓
landmark 3	(320,350,320)	(340,330,300)	✗

we exclude this record too.

The majority of vectors used to calculate correlation ρ is four, and the longest vector is 15. This is much shorter than the lengths of vectors used to calculate the correlation of residuals, so we may expect less precise estimation of $\rho(d)$.

Table 5.8 to Table 5.10 are the estimations of (5.30) for each landmark.

Table 5.8: Nonlinear least squares estimation of landmark 1.

	Estimate	Std. Error
c	0.372	0.062
λ	0.00091	0.00021

Table 5.9: Nonlinear least squares estimation of landmark 2.

	Estimate	Std. Error
c	0.184	0.054
λ	0.00076	0.00031

Table 5.10: Nonlinear least squares estimation of landmark 3.

	Estimate	Std. Error
c	0.250	0.080
λ	0.00161	0.00064

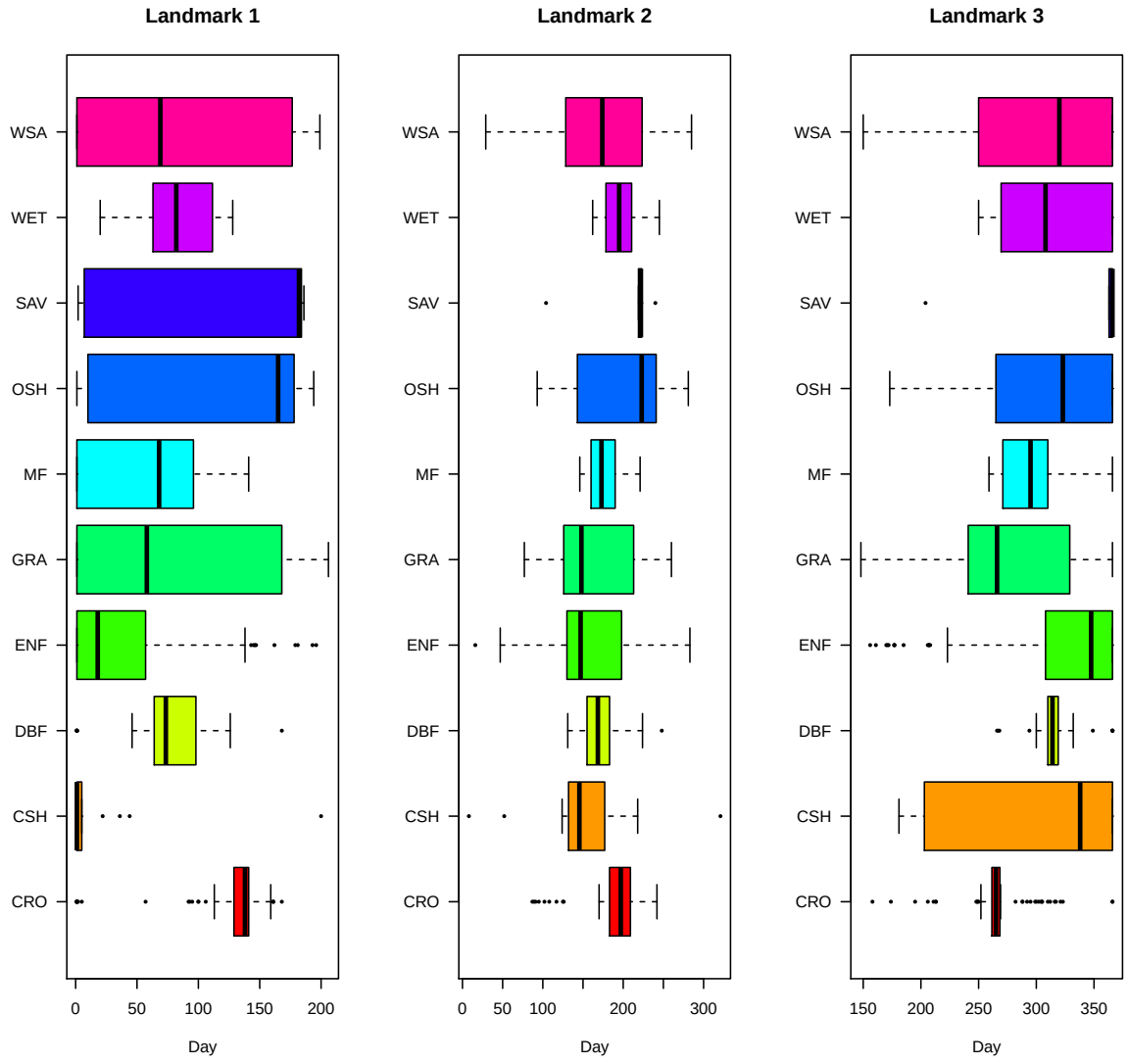


Figure 5.14: Landmarks against vegetation types. Each column presents how each landmark changes across vegetation types. The median of the second landmark has the least variation, while the median of the first landmark has the most variation. Each row displays how landmarks change with each vegetation type. For example, the third landmark of CSH has more variation than the other two.

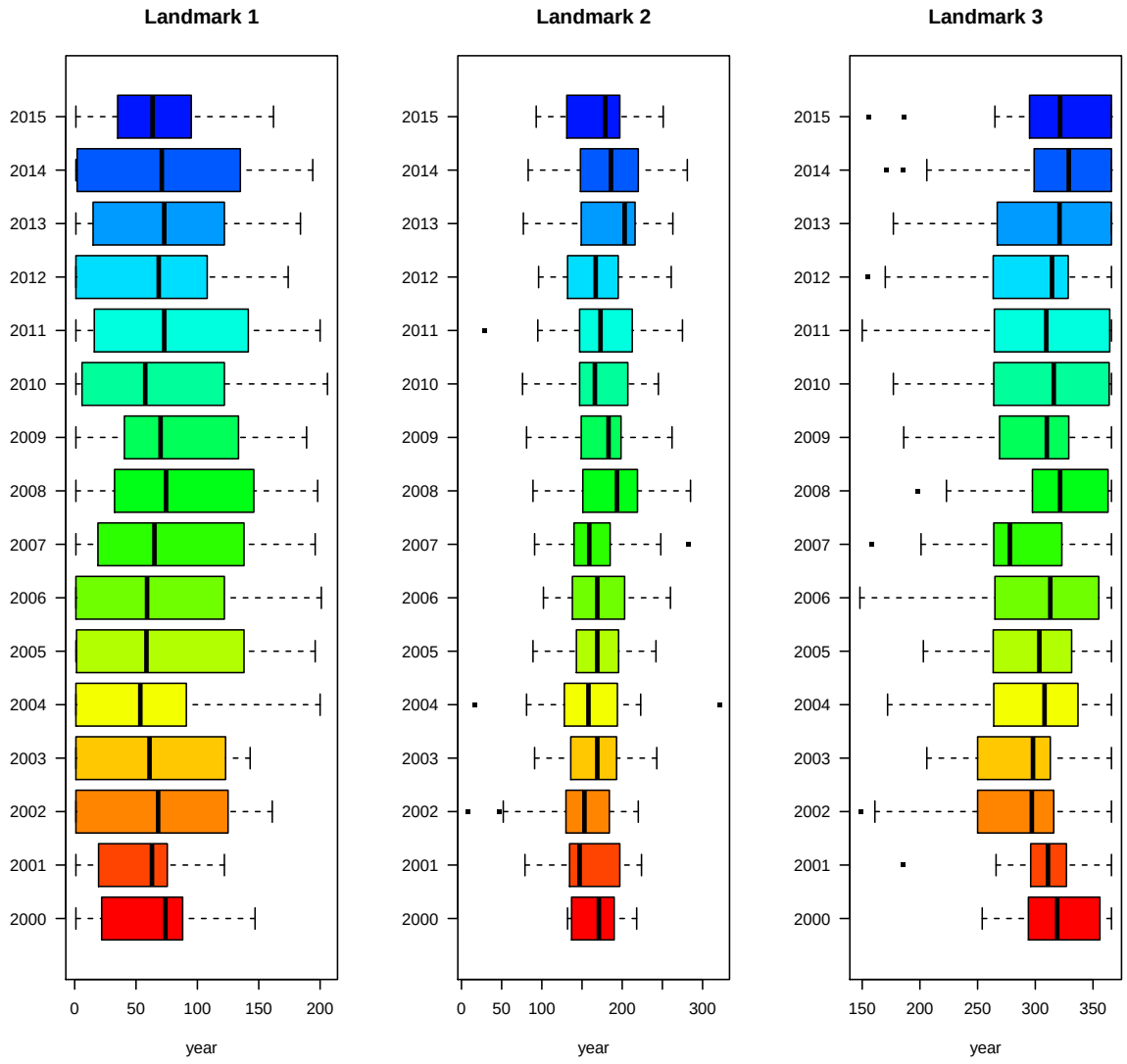


Figure 5.15: Landmarks against year. The median of three landmarks are stable across years. We have not spotted any trend for each landmark.

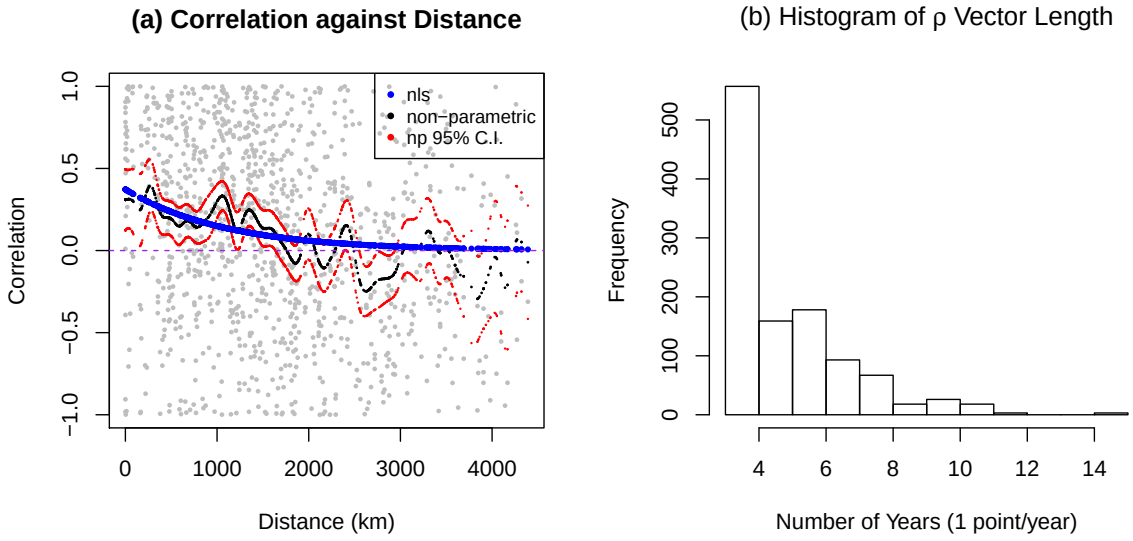


Figure 5.16: How landmark one's correlation changes with distance and the length of the vectors used to calculate the correlations. Image (a) displays the correlation against distance. The blue dots are the fitted exponential decay line, where the correlation decays to zero around 4000 *km*. The black dots are the fitting from non-parametric method and the red dots are the corresponding 95% confidence interval. Image (b) shows the histogram of the length of the vector used to calculate the correlation in years. The majority of pairs overlap in four years. We have only 1 point every year.

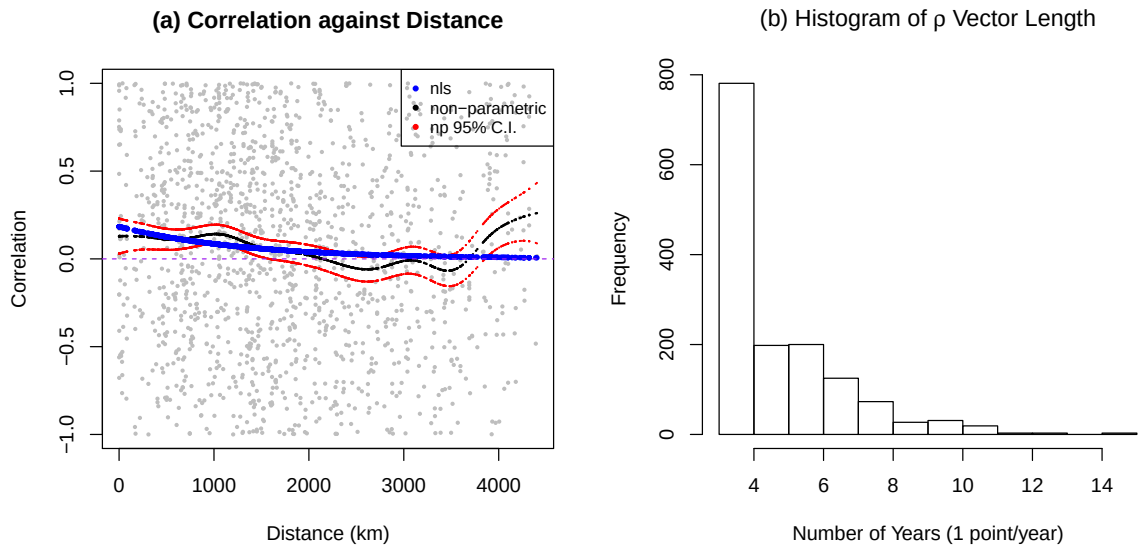


Figure 5.17: How landmark two's correlation changes with distance and the length of the vectors used to calculate the correlations. Image (a) displays the correlation against distance. The blue dots are the fitted exponential decay line, where the correlation decays to zero around 2500 *km*. The black dots are the fitting from non-parametric method and the red dots are the corresponding 95% confidence interval. Image (b) shows the histogram of the length of the vector used to calculate the correlation in years. The majority of pairs overlap in four years. We have only 1 point every year.

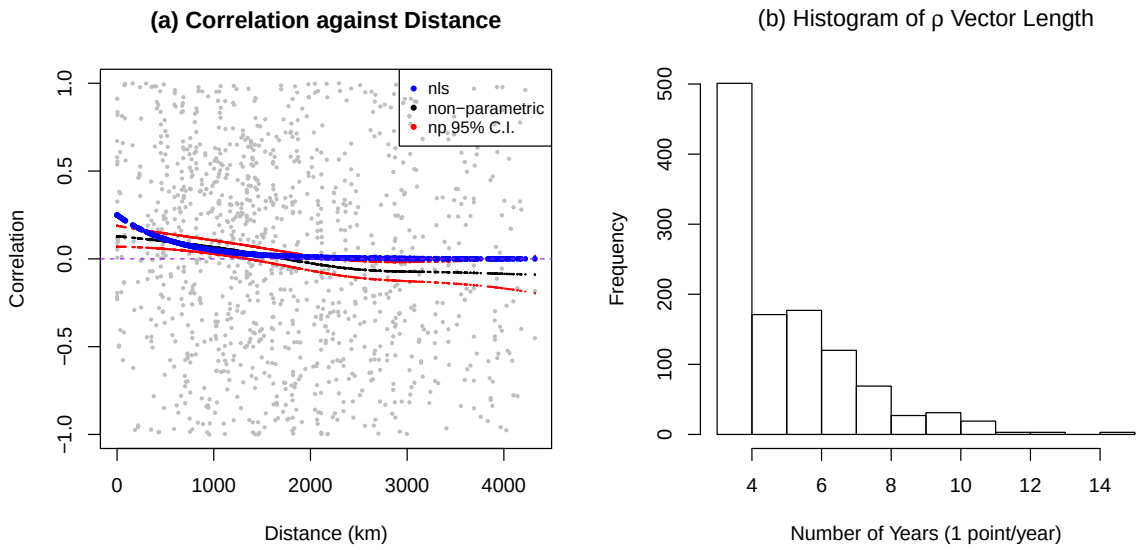


Figure 5.18: How landmark three's correlation changes with distance and the length of the vectors used to calculate the correlations. Image (a) displays the correlation against distance. The blue dots are the fitted exponential decay line, where the correlation decays to zero around 2000 *km*. The black dots are the fitting from non-parametric method and the red dots are the corresponding 95% confidence interval. Image (b) shows the histogram of the length of the vector used to calculate the correlation in years. The majority of pairs overlap in four years. We have only 1 point every year.

5.1.5 CEM summary

The CEM works based on the spatial correlation of each component. The spatial correlations are calculated based on data at the same time range. A big drawback of CEM is that for a location where the year we are trying to predict is out of the training dataset, we are unable to complete the mission. In our study, the training dataset ranges from 2002 to 2014, while the testing dataset ranges from 2002 to 2015. The sites of the testing dataset in 2015 are CA-TP4, US-SRG, US-SRM, US-Tw,1 US-Tw3, US-Twt, US-Whs, US-Wkg. Since we cannot have data for the same year in the training dataset, we are not able to come out a reasonable prediction for these sites. Another shortcoming is the spatial correlation of each component are all below 0.5, and only shows an effect within 1000 km.

For observed data, CEM updates the observation but does not update the underlying annual effect. For an unobserved site, this model theoretically doesn't work. We cheated a bit and calculated the MSE.

Our the assumptions of ϵ_{ijp} seems overly simplified. To model the spatial model, depending only on flux data may not be sufficient. We consider regression models in Sections 5.2 and 5.3 to overcome the problems.

5.2 Functional Linear Regression Model (FLRM)

The CEM approach was not very successful. Perhaps we need a different structured model, which not only uses the information of CO_2 flux but certain covariates as well. We considered using the NDVI data, introduced in Chapter 1 on page 5, to predict the CO_2 flux data. To achieve this goal we choose a functional regression model, which is a regression model with functional independent and dependent variables. We start by describing how we filtered these two types of data in Section 5.2.1. Section 5.2.2 briefly describes the model. The basis used in this model is shown in Section 5.2.3. Section 5.2.4 displays the results. Section 5.2.5 gives a short summary of FLRM.

5.2.1 Data Used in FLRM

For the 89 sites we selected in Section 5.1.1 we downloaded the NDVI data of each site at the same time coverage. The NDVI data were downloaded from the MODIS subsets tool on `https://modis.ornl.gov/sites/`. We selected the years when CO_2 flux and NDVI are both complete. After filtering there are 57 sites available. We plot them in Figure 5.19. Compared to Figure 5.2 we lost sites in the northern area.

Similarly to Figure 5.3, Figure 5.20 presents the year coverage, and number of sites each year. In CEM we have 89 sites with 536 years, but in the functional regression model, we only have 57 sites with 213 years. The summary of vegetation types of the 57 sites is shown in Table 5.11, whose distribution is still uneven. The number of ENF shrank the most, from 26 to 11, and GRA decreased from 17 to 10. For the two MF sites, both of them only have one year of data.

5.2.2 Modeling Details of FLRM

In the previous sections, we talked about the advantages of turning discrete data into functional objects. The functional regression model allows both independent variable and depen-

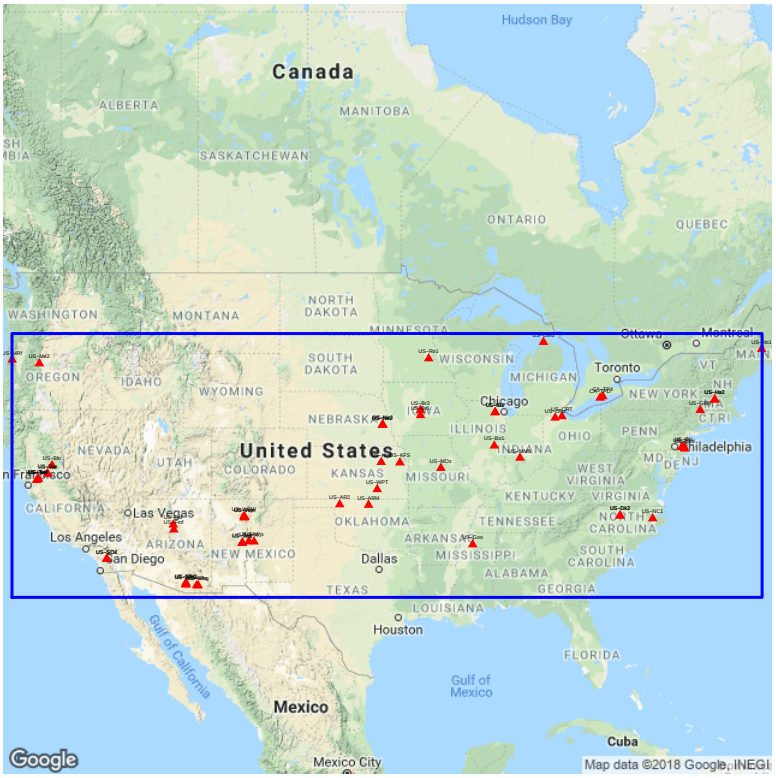


Figure 5.19: Locations of sites used in functional linear regression model. The squared area has latitude from 31°N to 46°N and longitude from 68.7°W to 123.6°W.

Table 5.11: Summary of number of sites for each species.

	Vegetation Type	**Number**
CRO	Croplands	12
CSH	Closed shrublands	3
DBF	Deciduous broadleaf forests	9
ENF	Evergreen needleleaf forest	11
GRA	Grasslands	10
MF	Mixed forest	2
OSH	Open shrublands	3
SAV	Savannas	1
WET	Permanent wetlands	3
WSA	Woody savannas	3
Total		57

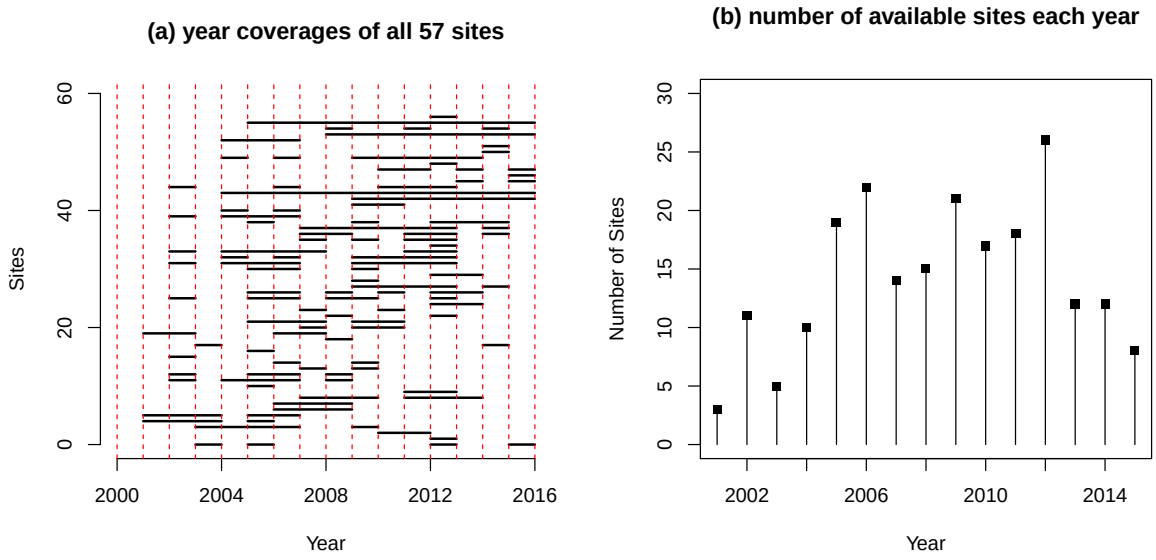


Figure 5.20: Year coverage of sites and number of sites each year. Image (a) plots the year coverage for each site. Image (b) summarize the number of sites in each year.

dent variable to be functional objects. Let y_{ip} be CO_2 flux data at site p of year i, z_{ip} be the corresponding NDVI data, and ϵ_{ip} be the measurement error. Their corresponding functional objects are $y_{ip}(t)$, $z_{ip}(t)$, and $\epsilon_{ip}(t)$. Since the processed CO_2 flux data and NDVI data have the same time-frequency, we use the concurrent model:

$$y_{ip}(t) = \alpha(t) + z_{ip}(t)\beta(t) + \epsilon_{ip}(t). \tag{5.60}$$

In this model the coefficients $\alpha(t)$ and $\beta(t)$ are also functional objects. We use function `fRegress` from R package `fda` to fit the model.

5.2.3 Basis Used in FLRM

The basis is cubic b-spline at knots (1,120,130,140,170,200,220,250,260,270,365).

5.2.4 Results of FLRM

We introduce the fitting results of FLRM in the following four sections: the correlation of flux(t) and NDVI(t) on day t, the training and testing datasets, the coefficients' estimations, and the predictions with the model assessment.

Data Correlation

We plot flux(t) against NDVI(t) on days $t = 1, 31, \ldots, 331$, in Figure 5.21. We used package `ggplot2` using "loess" method to add the smoothed red lines, which show the correlation between flux(t) and NDVI(t) changes over time. Days 91 to 241 in Figure 5.21 show a negative correlation between flux(t) against NDVI(t), while their correlations are not quite obvious during the other times. After calculating, the linear correlations of flux(t) and NDVI(t) on days $t = 91, 121, 151, 181, 211, 241$, are -0.315, -0.584, -0.66, -0.641, -0.595, and -0.456 respectively.

Training Data and Testing Data

We set the last year of each site as our testing data and the rest of the data as training data. For species MF, where each site has only one year of observations, it is not in the training dataset but the testing dataset. We show the complete CO_2 flux data and NDVI data in Figure 5.22 with their daily standard deviations. We can link the smoothed NDVI data in Table 1.3, which indicates the majority of NDVI data are collected from shrubs, grasslands, and rainforests. Though the annual pattern of NDVI is not as obvious as CO_2 flux, the trends of their daily standard deviation plots are similar. The peaks of the standard deviation of flux(t) and NDVI(t) are at $t = 190.8$ and $t = 192.6$. The training and testing datasets are plotted in Figures 5.23 and 5.24. The variations of training sets are similar to the overall sample, but the variation of NDVI testing set is different from the training set's.

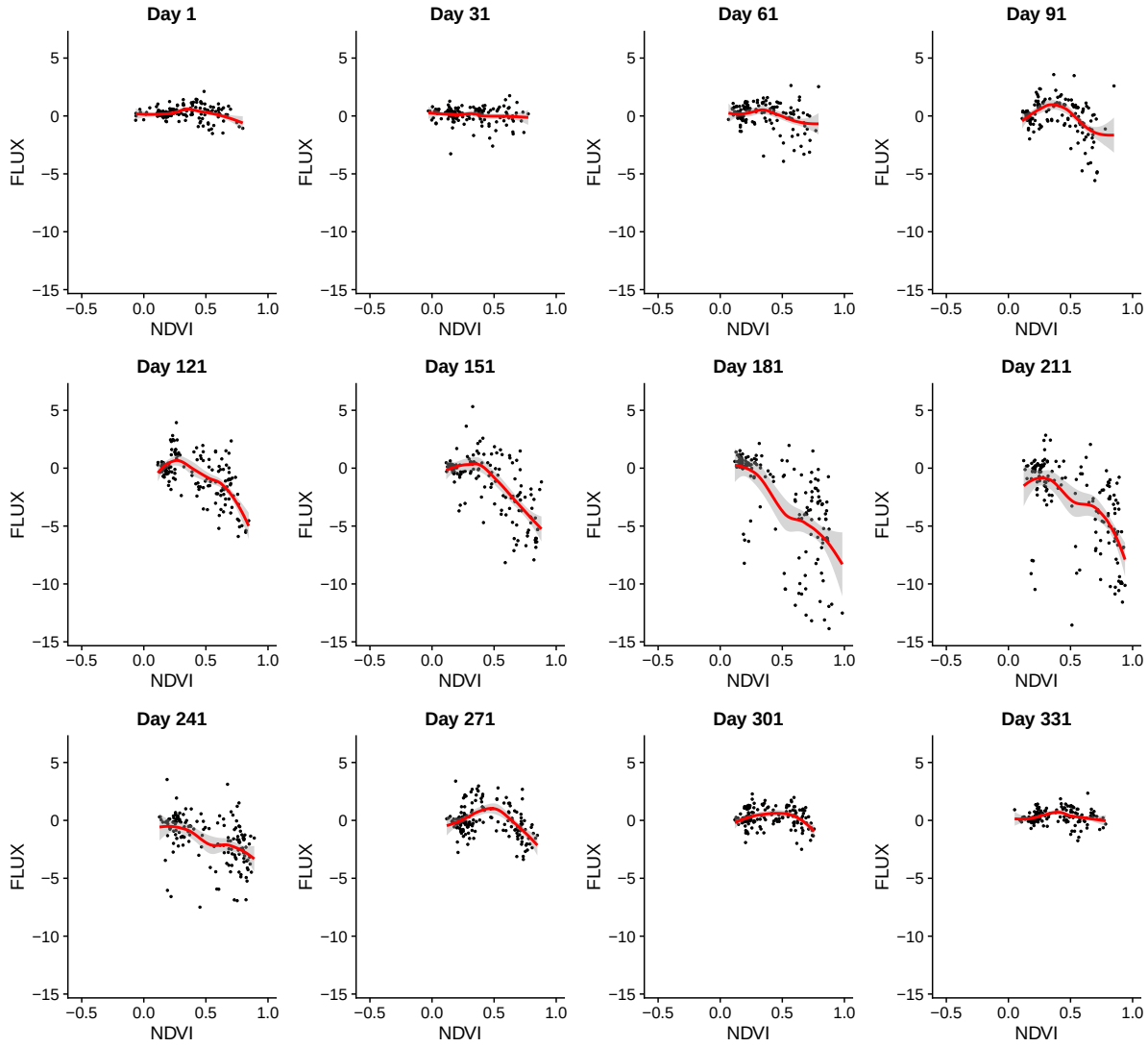


Figure 5.21: Plot flux(t) against NDVI(t) at every 30 days. The red lines are the estimated smoothed line using LOESS method. There is an obvious negative correlation between flux(t) and NDVI(t) for $t = 121, 151, 181, 211$.

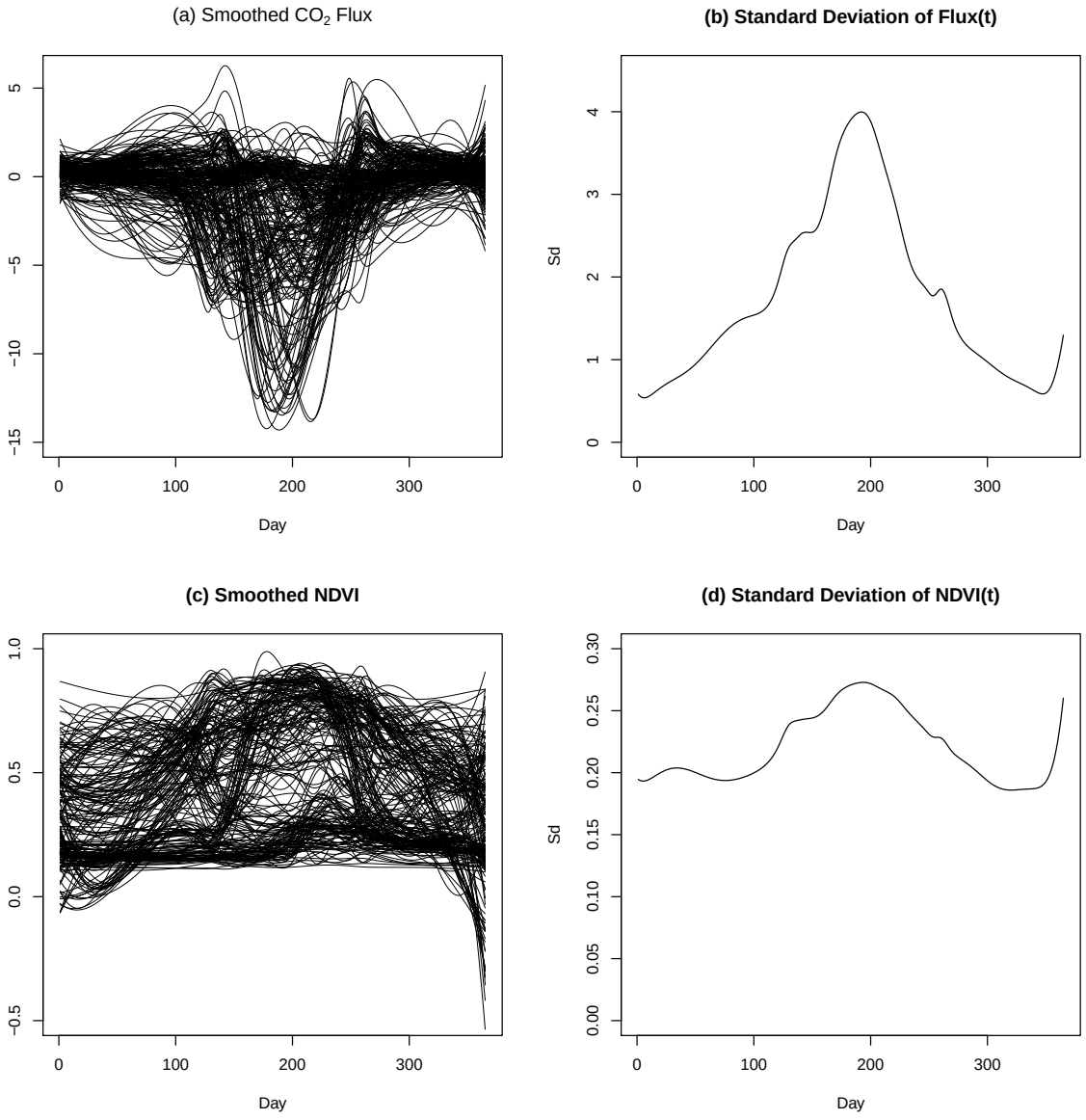


Figure 5.22: Data used in functional linear regression model. Image (a) plots the smoothed CO_2 flux data, whose point-wise standard deviation is plotted in image (b). Image (c) displays the smoothed NDVI, whose point-wise standard deviation is plotted in image (d).

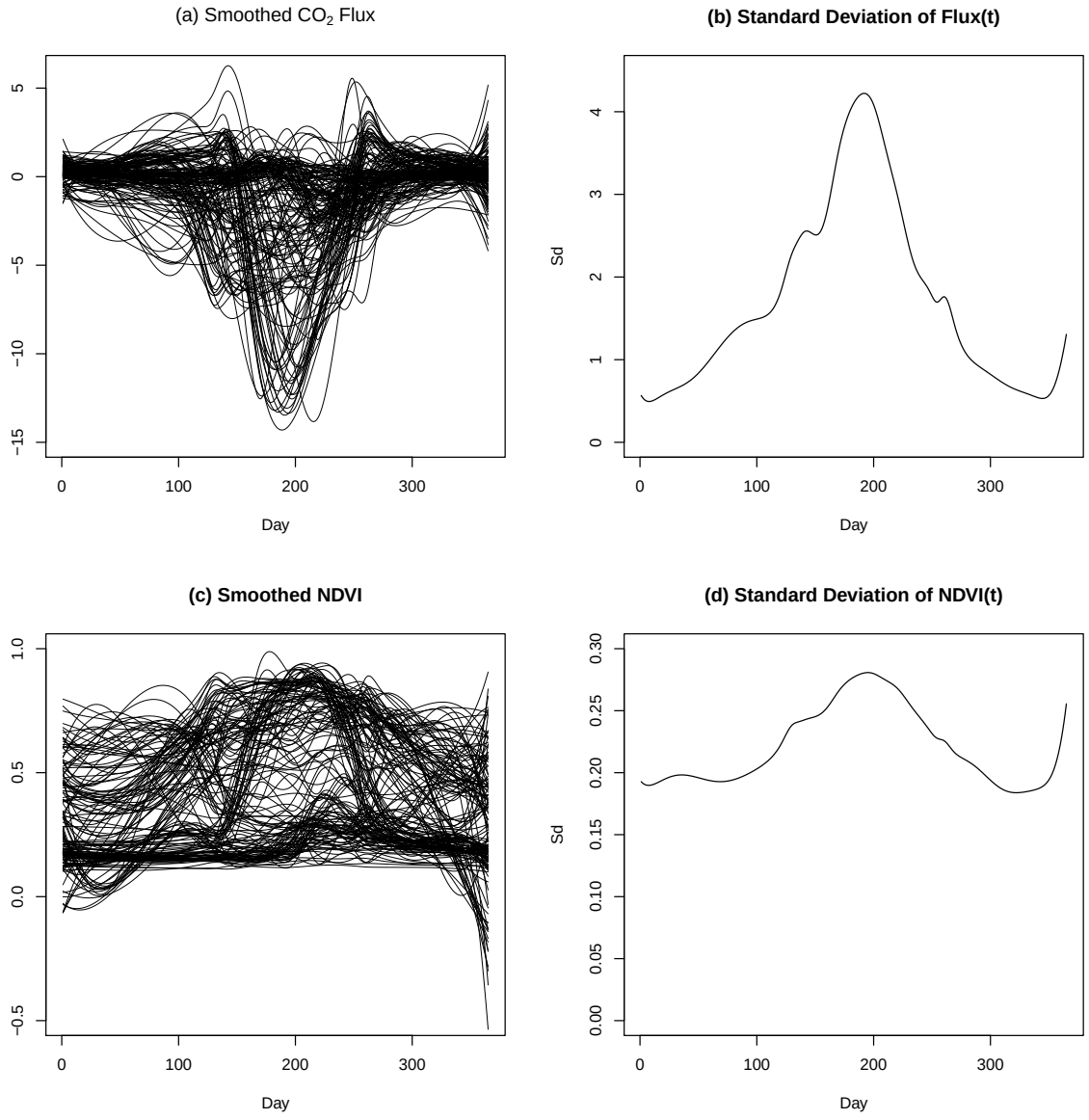


Figure 5.23: Training data used in functional linear regression model. Image (a) plots the smoothed CO_2 flux data, whose point-wise standard deviation is plotted in image (b). Image (c) displays the smoothed NDVI, whose point-wise standard deviation is plotted in image (d).

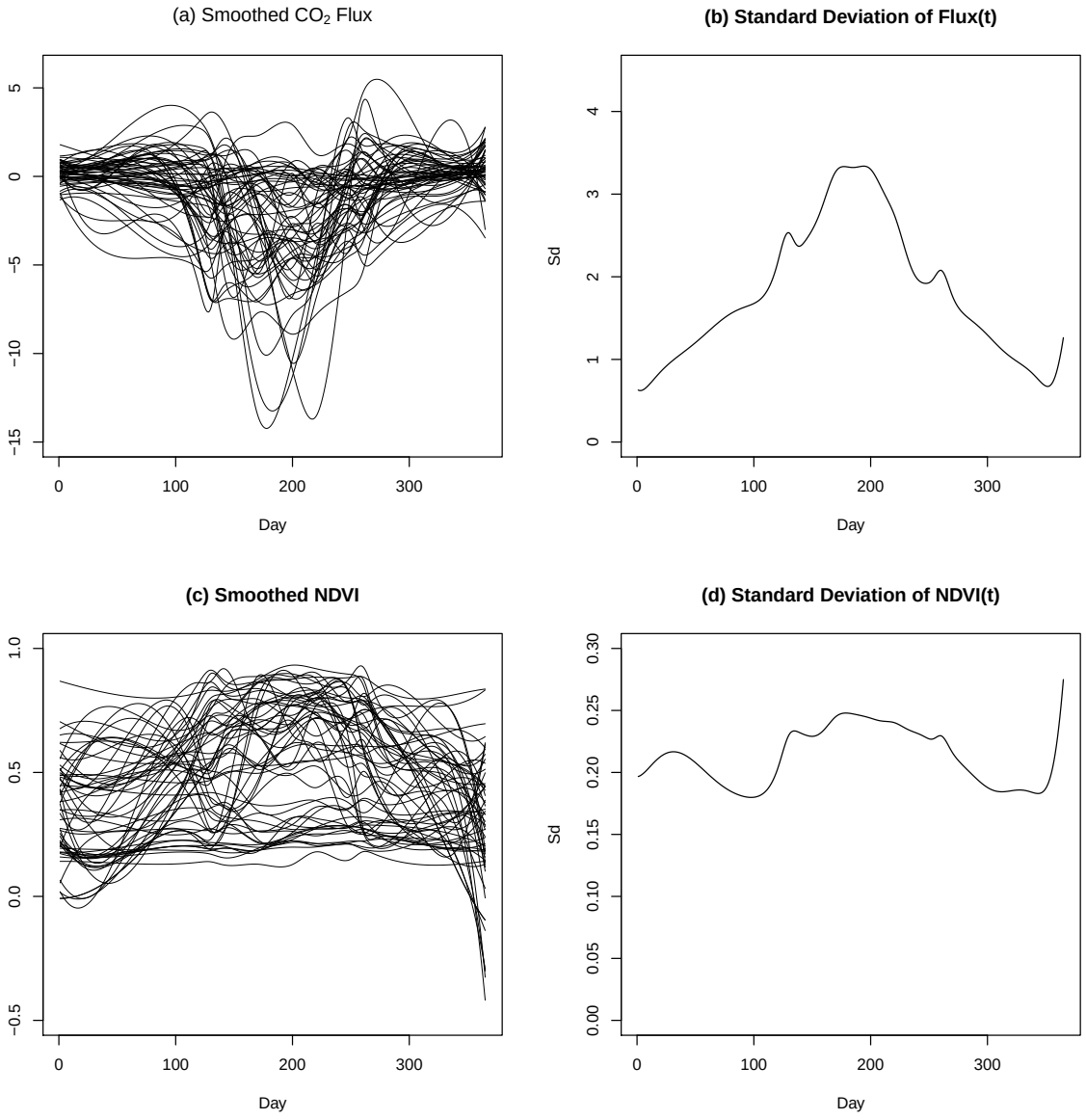


Figure 5.24: Testing data used in functional linear regression model. Image (a) plots the smoothed CO_2 flux data, whose point-wise standard deviation is plotted in image (b). Image (c) displays the smoothed NDVI, whose point-wise standard deviation is plotted in image (d).

Estimations of Coefficients

The coefficients estimations of $\alpha(t)$ and $\beta(t)$ are shown in Figure 5.25. The $\hat{\alpha}(t)$ and $\hat{\beta}(t)$ have opposite effects. To further view the effects of $\hat{\alpha}(t)$ and $\hat{\beta}(t)$, we plot them together with predictions in Figure 5.26.

Predictions

Figure 5.26 presents the prediction of the first six sites, and the two MF sites. FLRM didn't capture the trend of site US-ARM and US-GMF. For US-ARM, the observations are above $\hat{\alpha}(t)$, $t \in [150, 270]$. For this period, the NDVI is usually positive and reach the peak of the year, which means the model has difficulties to predict positive CO_2 flux. In the last row of Figure 5.26, though the vegetation type of these two sites is not in the training dataset, the model captured the "V" shaped trend in US-Dix.

5.2.5 FLRM Summary

Though the FLRM used only one covariate, it captures the trend of CO_2 flux. This model works based on the correlation between CO_2 flux and NDVI. In our study, we show the correlation change over time in Figure 5.21. Overall, the covariate we picked has a negative correlation with CO_2 flux.

Inspired by the Canadian weather and kneecap studies, to further improve this model we can select and use the covariate with higher correlation with CO_2 flux, or apply the non-concurrent model. Combining more covariates in the model, such as location, elevation, vegetation type is another way.

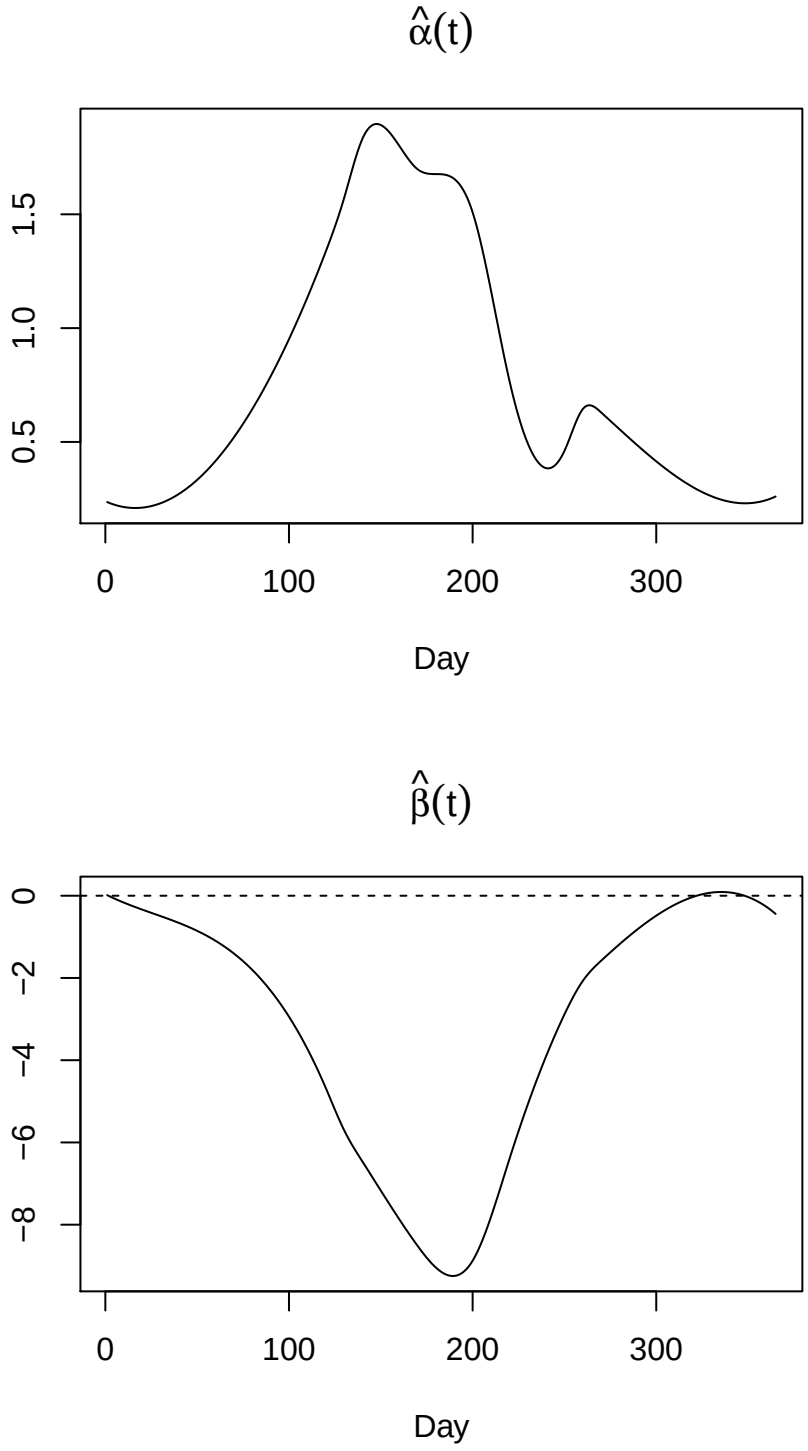


Figure 5.25: Coefficients estimations of functional linear regression model (5.60). Image (a) plots the estimation of the $\alpha(t)$. Image (b) displays the estimation of $\beta(t)$.

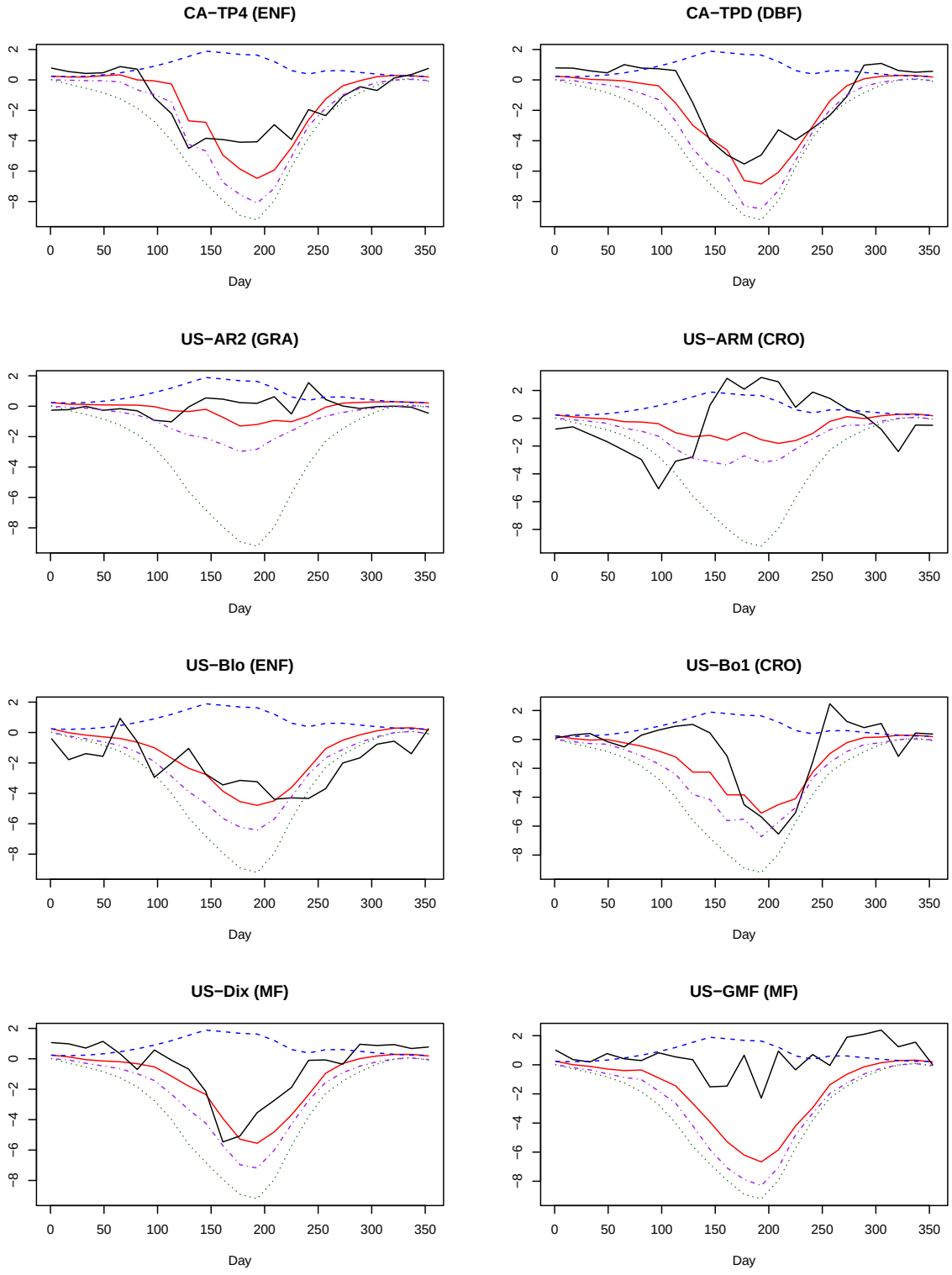


Figure 5.26: Predictions functional linear regression model using smoothed NDVI data. The black solid lines are observations, the red solid lines are predictions, the dashed blue lines are $\hat{\alpha}(t)$, the dashed purple lines are $z_{ip}(t)\hat{\beta}(t)$, and the dashed green lines are $\hat{\beta}(t)$.

5.3 Generalized Additive Model (GAM)

In the FLRM, we only used NDVI as the covariate. There is other information available, such as location (longitude and latitude), vegetation type, and elevation. Under the FLRM structure, it's not easy to embed other information.

Instead of treating the flux data as repeated annual data with a similar trend, we consider it as a single time series. GAM is a powerful tool to fit a model with flexible combinations of the covariates. In Section 5.3.1, we briefly introduce the data used in GAM. Section 5.3.2 describes the selected model. The basis used for each smoothed term is introduced in Section 5.3.3. The results are shown in Section 5.3.4.

5.3.1 Data used in GAM

We use the same training data in GAM as FLRM.

5.3.2 Details of GAM

Beyond using NDVI to predict flux data, we are also curious about whether using more information, such as latitude, longitude, elevation, vegetation, and NDVI, generates a better model.

After several experiments, we chose the model with minimal GCV:

$$\texttt{flux} \sim f(\texttt{latitude}, \texttt{longitude}) + \texttt{vegetation} + f(\texttt{elevation}, \texttt{NDVI}). \tag{5.61}$$

5.3.3 Basis of GAM

According to Wood (2006), the interaction of latitude and longitude, because they are measured in the same units, we use thin plate regression splines. As for elevation and NDVI, which are measured in different units, we choose the tensor product.

5.3.4 Results of GAM

We show part of the predictions in Figure 5.27.

5.3.5 GAM Summary

GAM has a lot more flexibility of adding covariates than FLRM. All effects are additive, and it's easier to visualize each effect. Since more information is involved in GAM, how to properly arrange the covariates is one of the challenges. One drawback of our GAM is that it cannot predict for species that are not in the training model, for example, MF. Though more information is used in GAM, for this type of data, the predictions are not dramatically improved. The MSE indicates GAM is better than FLRM.

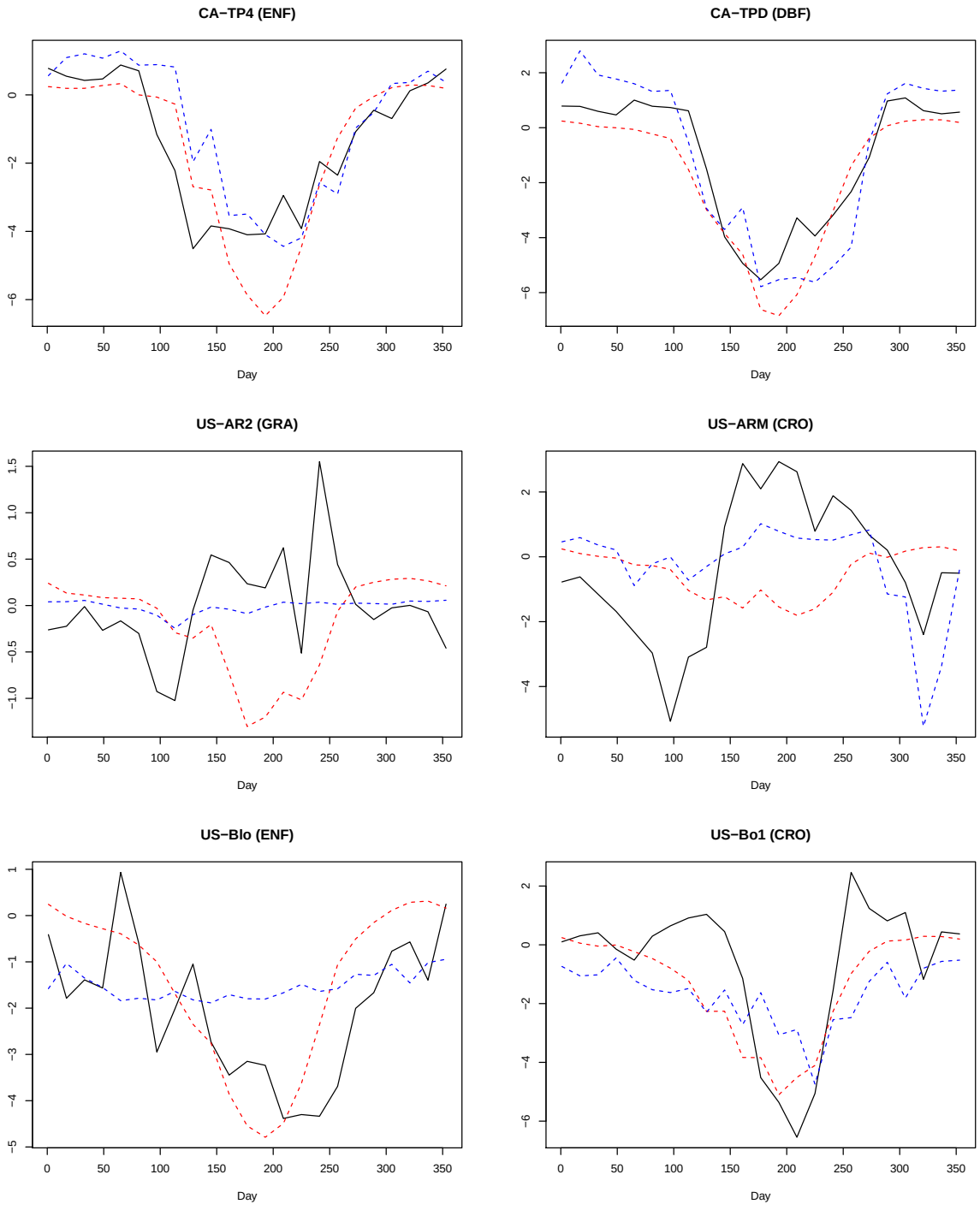


Figure 5.27: Model prediction comparison. The black solid lines are observations, the blue dashed lines are predictions from GAM, the red dashed lines are predictions from FLRM. For sites CA-TP4, CA-TPD, and US-Bo1, the predictions from GAM and FLRM have the same trend.For sites US-AR2, and US-Blo, the predictions from GAM are more flat, while the predictions from FLRM are more V shaped.For site US-ARM, the predictions from neither models captured the overall trend.

5.4 Discussion

We present the MSE of the three models on the same testing dataset in Table 5.12.

Table 5.12: MSE of three spatial models in units of $(\mu mol\ CO_2\ m^{-2}s^{-1})^2$.

CEM	FLRM	GAM
4.78	3.25	2.91

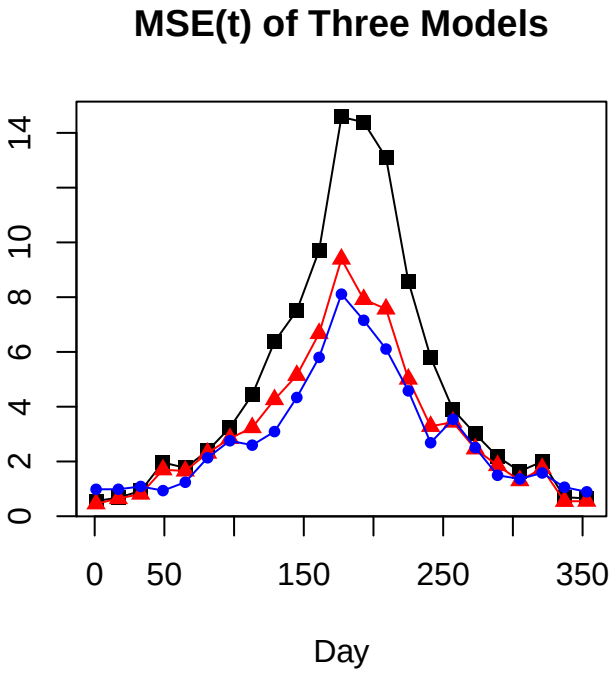


Figure 5.28: MSE(t) of three spatial models on the same testing dataset. (The black squares and lines are from CEM, the red triangles and lines are predictions from FLRM, and the blue points and lines are from GAM.)

We left the species out from the GAM model and then predicted for MF in Figure 5.29. For predictions of MF, FLRM is better than GAM.

For our data, the GAM achieves the lowest MSE. Although the FLRM uses much less information than GAM, FLRM's MSE is slightly bigger than GAM's.

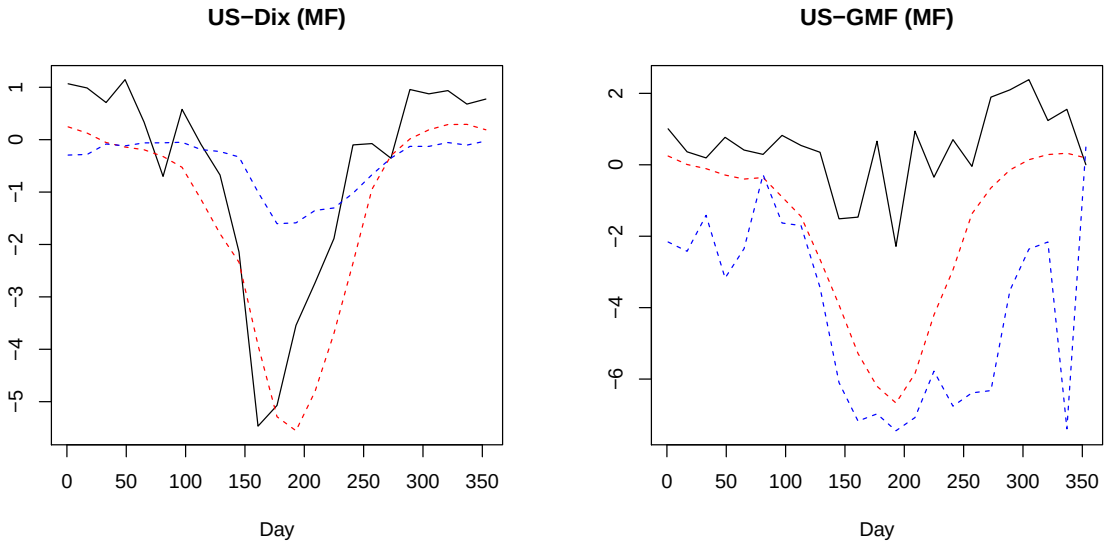


Figure 5.29: Model prediction comparison. The black solid lines are observations, the blue dashed lines are predictions from GAM, the red dashed lines are predictions from FLRM.

Chapter 6

Conclusion

Certain types of data, like CO_2 flux, are challenging because they are continuous in time, repeated in patterns, and vary according to phase. In Wood (2006), the Chicago air pollution case gives another example of ecological data. Other examples would be the environmental data, medical data, etc. The problems associated with these data are

1. predicting future observations. One of the challenges is to properly model phase variation.
2. comparing data over time rather than at limited time points.

We can study the data in both discrete and continuous forms. The advantages of the continuous form are the ability to capture the underlying smooth process of the data and not be distracted by the measurement noise. The continuous form allows us to evaluate the value at any point within the domain. This allows us to compare data over continuous time intervals rather than limited time points. Because of the continuity, more information can be retrieved from the derivatives.

Compared to interpolation, which goes through each data point, smoothing techniques reduce measurement error, though it cannot be completely eliminated from experiments.

In this thesis we addressed these issues in two ways: single location model and spatial models. We summarize the contribution of our single and spatial models in Sections 6.1 and

6.2 respectively. In Section 6.3, we list potential topics for future study.

6.1 Single Location Modeling

For the single location study, the problems we addressed were modeling phase variation and generating reasonable predictions. We can either treat the observations as a single time series modeled through GAM or take advantage of the data's replication using FDA. Because our GAM assumed the underlying annual cycles are the same, it did not help with the phase variation problem. If we ignore the effect of seasonal dynamics, the vertical variation is inflated as shown in Figure 4.19. Our single location model chooses a "detour" strategy using registration, because registration improves the model fit. We first smooth the data, next we use landmark registration to diminish the variation caused by phase variation, then we study the features of landmarks. In the end, we map the data back to the real time scale. For the landmark study, both Dirichlet regression and CDA can be used. We chose Dirichlet regression because its parameters are more interpretable. We believe this is the first application of these techniques in this context.

6.2 Spatial Modeling

For spatial modeling, we used three models:

1. CEM. This model uses only CO_2 flux data to predict flux at various unobserved sites. We assume the CO_2 flux data are made up of different components, which are the smoothed overall effect, the site effect, year effect and measurement error. We calculated the spatial correlation of each component to fit the exponential decay model $\rho(d) = c\exp(-\lambda d)$. Unfortunately, the spatial correlations of all components are below 0.5 and only provide useful predictions for sites within 1000 *km* distance. By means of conditional expectation, theoretically, we should be able to predict CO_2 flux for the site of interest, but there was little variation among the predictions. This model failed to capture the trends

of different vegetation types.

2. FLRM. After noticing that CO_2 flux data alone may not be sufficient to build a decent spatial model, we chose NDVI data as the covariate because the NDVI data is correlated with the CO_2 flux data and freely available everywhere. Unlike a simple linear regression model, in FLRM, we treat both NDVI and CO_2 flux data as functional objects. For vegetation types CRO and GRA, the data patterns are influenced by natural causes but also by anthropic factors, and so they are sometimes not consistent with other types. If we embed other covariates, this may improve the model. But this is not easy for FLRM, because of problems related to both computation and interpretation.
3. GAM. We used GAM because of its flexibility in adding covariates, which is a big advantage over FLRM. The covariates we used were latitude, longitude, elevation, vegetation types, and NDVI. The GAM model has the smallest MSE among the three models.

Our report of our experience with these models is the main contribution of this chapter.

6.3 Future Work

Our study provides a basis for future research in several areas, which include:

1. In single location modeling, we used only the numeric method to validate the model. The theoretical approach can help to reduce the hours of computation and be applied universally. To achieve this, we need a series of approximations of each step, so we may carefully calculate and control the errors in each step.
2. We used Dirichlet regression to model the landmarks. Through the coefficients, we can calculate the range of each phase. However, it would be more beneficial in predicting the direction or trend of each phase change. This task is not feasible for our current dataset because the length of observation each site is not long enough. Consequently, we may look for other related data sources.
3. The registration idea works well for the single location study. We may merge it into the

FLRM. The idea is to build an FLRM using registered CO_2 flux and NDVI data. When given NDVI, we would first register the NDVI and generate predictions from the model. Then we would model the landmarks and map the predictions back to the real time scale. In this way, we may reduce prediction error.

4. We are interested in calculating the minimum numbers of years of data required for such studies. The solution to this question involves proper study design and power calculation.

Bibliography

A. P. Camargo, J. M. Stern, M. S. Lauretto, P. Goyal, A. Giffin, K. H. Knuth, and E. Vrscay. Estimation and model selection in dirichlet regression. In AIP Conference Proceedings-American Institute of Physics, volume 1443, page 206, 2012.

H. A. Cleugh, R. Leuning, Q. Mu, and S. W. Running. Regional evaporation estimates from flux tower and MODIS satellite data. Remote Sensing of Environment, 106(3):285–304, 2007.

N. Cressie. Mission co2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide. Journal of the American Statistical Association, 113(521):152–168, 2018.

A. C. de Araújo. The Brazilian scientist Antonio Donato Nobre from the INPA, Manaus, contemplating the futute and functioning of the amazon from the top of the micrometeorological flux tower (54 m) in the cuieiras reserve, 100 km north of manaus, part of CARBONSINK-LBA. Retrieved from: `http://carboeurope.org/ceip/products/foto.html`, Nov. 2015. Accessed: 2016-09-16.

B. Efron. Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7(1):1–26, 1979.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35(3):279–300, 2003.

I. Epifanio and N. Ventura-Campos. Functional data analysis in shape analysis. Computational Statistics & Data Analysis, 55(9):2758–2773, 2011.

I. Epifanio and N. Ventura-Campos. Hippocampal shape analysis in alzheimer's disease using functional data analysis. Statistics in Medicine, 33(5):867–880, 2014.

FLUXNET. Distribution of tower sites within the global network of networks. Retrieved from: https://fluxnet.ornl.gov/introduction, 2014. Accessed: 2016-09-16.

C. Gorrostieta, J. Ortega, A. J. Quiroz, and G. H. Smith. Characterization of storm wave asymmetries with functional data analysis. Environmental and Ecological Statistics, 21(2): 263–283, 2014.

T. Hastie and R. Tibshirani. Generalized Additive Models. Wiley Online Library, 1990.

T. Hayfield, J. S. Racine, et al. Nonparametric econometrics: The np package. Journal of Statistical Software, 27(5):1–32, 2008.

F. A. Heinsch, M. Zhao, S. W. Running, J. S. Kimball, R. R. Nemani, K. J. Davis, P. V. Bolstad, B. D. Cook, A. R. Desai, D. M. Ricciuto, et al. Evaluation of remote sensing based terrestrial productivity from MODIS using regional tower eddy flux network observations. IEEE Transactions on Geoscience and Remote Sensing, 7(44):1908–1925, 2006.

R. H. Hijazi and R. W. Jernigan. Modelling compositional data using dirichlet regression models. Journal of Applied Probability & Statistics, 4(1):77–91, 2009.

K. R. Hogrefe, V. P. Patil, D. R. Ruthrauff, B. W. Meixell, M. E. Budde, J. W. Hupp, and D. H. Ward. Normalized difference vegetation index as an estimator for abundance and quality of avian herbivore forage in arctic alaska. Remote Sensing, 9(12):1234, 2017.

J. L. Kalfas, X. Xiao, D. X. Vanegas, S. B. Verma, and A. E. Suyker. Modeling gross primary production of irrigated and rain-fed maize using MODIS imagery and CO2 flux tower data. Agricultural and Forest Meteorology, 151(12):1514–1528, 2011.

M. D. King, C. L. Parkinson, K. C. Partington, and R. G. Williams. Our Changing Planet: the View from Space. Cambridge University Press New York, New York, USA, 2007.

S. Klosterman, K. Hufkens, J. Gray, E. Melaas, O. Sonnentag, I. Lavine, L. Mitchell, R. Norman, M. Friedl, and A. Richardson. Evaluating remote sensing of deciduous forest phenology at multiple spatial scales using phenocam imagery. Biogeosciences Discuss, page 43054320, 2014.

B. Krasean. CO2 flux towers help assess the sustainability of biofuels. Retrieved from: http://lter.kbs.msu.edu/2013/08/co2-flux-towers-help-assess-the-sustainability-of-biofuels/, 2013. Accessed: 2016-09-16.

M. E. Krasny. Invasion ecology. NSTA Press, 2003.

C. Kuenzer, S. Dech, and W. Wagner. Remote sensing time series revealing land surface dynamics: Status quo and the pathway ahead. In Remote Sensing Time Series, pages 1–24. Springer, 2015.

X. Leng and H.-G. Müller. Classification using functional data analysis for temporal gene expression data. Bioinformatics, 22(1):68–76, 2006.

LI-COR. 6400-09 Soil CO2 flux chamber. Retrieved from: https://www.licor.com/env/products/photosynthesis/LI-6400XT/chambers/6400-09.html, 2016. Accessed: 2016-09-17.

M. J. Maier. Dirichletreg: Dirichlet regression for compositional data in r. Retrieved from: https://cran.r-project.org/web/packages/DirichletReg/vignettes/DirichletReg-vig.pdf, 2014. Accessed: 2016-09-16.

M. J. Maier. DirichletReg: Dirichlet Regression in R, 2015. URL http://dirichletreg.r-forge.r-project.org/. R package version 0.6-3.

A. M. Moffat, D. Papale, M. Reichstein, D. Y. Hollinger, A. D. Richardson, A. G. Barr, C. Beckstein, B. H. Braswell, G. Churkina, A. R. Desai, et al. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agricultural and Forest Meteorology, 147(3):209–232, 2007.

NASA. NASA's earth observing system: Project science office. Retrieved from: http://eospso.nasa.gov/, 2016. Accessed: 2016-09-16.

NOAA. Polar orbiting: Aqua Satellite and MODIS Swath. Retrieved from: http://sos.noaa.gov/Datasets/dataset.php?id=34, 2016. Accessed: 2016-09-16.

C. Parkinson, A. Ward, and M. King. Earth science reference handbook: A guide to nasa's earth science program and earth observing satellite missions. National Aeronautics and Space Administration, page 277, 2006.

V. Pawlowsky-Glahn and A. Buccianti. Compositional Data Analysis: Theory and Applications. John Wiley & Sons, 2011.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org.

J. O. Ramsay and X. Li. Curve registration. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(2):351–363, 1998.

J. O. Ramsay, H. Wickham, S. Graves, and G. Hooker. Fda: Functional Data Analysis, 2018. URL https://CRAN.R-project.org/package=fda. R package version 2.4.8.

T. S. Tian. Functional data analysis in brain imaging studies. Frontiers in Psychology, 1, 2010.

K. G. Van den Boogaart and R. Tolosana-Delgado. Analyzing Compositional Data with R, volume 122. Springer, 2013.

W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

J. Weier and D. Herring. Measuring vegetation (ndvi & evi), September 2018. URL `https://earthobservatory.nasa.gov/Features/MeasuringVegetation`. [Online; posted 30-August-2000].

S. N. Wood. Mgcv: Gams and generalized ridge regression for r. R News, 1(2):20–25, 2001.

S. N. Wood. Generalized Additive Models: an Introduction with R. CRC press, 2006.

F. Yang, M. A. White, A. R. Michaelis, K. Ichii, H. Hashimoto, P. Votava, A.-X. Zhu, and R. R. Nemani. Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. IEEE Transactions on Geoscience and Remote Sensing, 44(11):3452–3461, 2006.

F. Yang, K. Ichii, M. A. White, H. Hashimoto, A. R. Michaelis, P. Votava, A.-X. Zhu, A. Huete, S. W. Running, and R. R. Nemani. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. Remote Sensing of Environment, 110(1):109–122, 2007.

Appendix A

Additional Tables

Table A.1: 18 sites of CCP that have CO_2 flux above-canopy data.

Site Name	Data Collection Timeframe
BC-Campbell River 1949 Douglas-fir	1997-2015
BC-Campbell River 1988 Douglas-fir	2001-2015
BC-Campbell River 2000 Douglas-fir	2000-2015
NB-Charlie Lake 1 1975 Balsam fir	2003-2005
ON-Borden Mixedwood	1995-2015
ON-Groundhog River Mixedwood	2003-2015
ON-Turkey Point 1939 White Pine	2003-2015
ON-Turkey Point 1974 White Pine	2003-2015
ON-Turkey Point 1989 White Pine	2003-2015
ON-Turkey Point Deciduous	2012-2014
QC-2000 Harvested Black Spruce/Jack Pine (HBS00)	2001-2010
QC-Eastern Old Black Spruce (EOBS)	2003-2010
SK-1975 Young Jack Pine	2003-2015
SK-2002 Jack Pine	2003-2015
SK-1994 Jack Pine	2001-2015
SK-Old Aspen	1996-2015
SK-Old Jack Pine	1994-2015
SK-Southern Old Black Spruce	1994-2015

Data collected at BC-Campbell River area had only one species, Douglas fir. The NB-Carlie Lake did not provide sufficient data for this study. As for ON-Turkey Point, the information of exact canopy height was not available. ON-Borden Mixedwood data included more than one species.

Table A.2: MODIS products.

MOD	01	Level-1A Radiance Counts
MOD	02	Level-1B Calibrated Geolocation Data Set
MOD	03	Geolocation Data Set
MOD	04	Aerosol Product
MOD	05	Total Precipitable Water
MOD	06	Cloud Product
MOD	07	Atmospheric Profiles
MOD	08	Gridded Atmospheric Product
MOD	09	Surface Reflectance; Atmospheric Correction Algorithm Products
MOD	10	Snow Cover
MOD	11	Land Surface Temperature and Emissivity
MOD	12	Land Cover/Land Cover Change
MOD	13	Gridded Vegetation Indices (NDVI & EVI)
MOD	14	Thermal Anomalies - Fires and Biomass Burning
MOD	15	Leaf Area Index (LAI) and Fractional Photosynthetically Active Radiation (FPAR)
MOD	16	Evapotranspiration
MOD	17	Vegetation Production, Net Primary Productivity (NPP)
MOD	18	Normalized Water - leaving Radiance
MOD	19	Pigment Concentration
MOD	20	ChlorophyII Fluorescence
MOD	21	ChlorophyII_a Pigment Concentration
MOD	22	Photosynthetically Available Radiation (PAR)
MOD	23	Suspended - Solids Concentration
MOD	24	Organic Matter Concentration
MOD	25	Coccolith Concentration
MOD	26	Ocean Water Attenuation Coefficient
MOD	27	Ocean Primary Productivity
MOD	28	Sea Surface Temperature
MOD	29	Sea Ice Cover
MOD	31	Phycoerythrin Concentration
MOD	32	Processing Framework and Match-up Database
MOD	35	Cloud Mask
MOD	36	Total Absorption Coefficient
MOD	37	Ocean Aerosol Properties
MOD	39	Clean Water Epsilon
MOD	40	Gridded Thermal Anomalies
MOD	43	Surface Reflectance BRDF/Albedo Parameter
MOD	44	Vegetation Cover Conversion

Table A.3: The selected 89 AmeriFlux Sites.

	Site	Latitude	Longitude	Elevation (m)	Vegetation Type
1	CA-Gro	48.2167	-82.1556	340	MF
2	CA-NS3	55.9117	-98.3822	260	ENF
3	CA-NS6	55.9167	-98.9644	244	OSH
4	CA-Qcu	49.2671	-74.0365	392	ENF
5	CA-Qfo	49.6925	-74.3421	382	ENF
6	CA-TP1	42.6609	-80.5595	265	ENF
7	CA-TP3	42.7068	-80.3483	184	ENF
8	CA-TP4	42.7102	-80.3574	184	ENF
9	CA-TPD	42.6353	-80.5577	260	DBF
10	US-AR1	36.4267	-99.4200	611	GRA
11	US-AR2	36.6358	-99.5975	646	GRA
12	US-ARM	36.6058	-97.4888	314	CRO
13	US-Bar	44.0646	-71.2881	272	DBF
14	US-Bkg	44.3453	-96.8362	510	GRA
15	US-Blk	44.1580	-103.6500	1718	ENF
16	US-Blo	38.8953	-120.6328	1315	ENF
17	US-Bo1	40.0062	-88.2904	219	CRO
18	US-Br1	41.6915	-93.6914	313	CRO
19	US-Br3	41.9747	-93.6936	313	CRO
20	US-Ced	39.8379	-74.3791	58	CSH
21	US-CRT	41.6285	-83.3471	180	CRO
22	US-Dix	39.9712	-74.4346	48	MF
23	US-Dk1	35.9712	-79.0934	168	GRA
24	US-Dk2	35.9736	-79.1004	168	DBF
25	US-Dk3	35.9782	-79.0942	163	ENF
26	US-Fmf	35.1426	-111.7273	2160	ENF
27	US-FPe	48.3077	-105.1019	634	GRA
28	US-Fwf	35.4454	-111.7718	2270	GRA
29	US-GLE	41.3665	-106.2399	3197	ENF
30	US-GMF	41.9667	-73.2333	380	MF
31	US-Goo	34.2547	-89.8735	87	GRA
32	US-Ha1	42.5378	-72.1715	340	DBF
33	US-Ha2	42.5393	-72.1779	360	ENF
34	US-Ho1	45.2041	-68.7402	60	ENF
35	US-Ho2	45.2091	-68.7470	91	ENF
36	US-Ho3	45.2072	-68.7250	61	ENF
37	US-IB1	41.8593	-88.2227	226	CRO
38	US-IB2	41.8406	-88.2410	226	GRA
39	US-KFS	39.0561	-95.1907	310	GRA
40	US-Kon	39.0824	-96.5603	330	GRA
41	US-KUT	44.9950	-93.1863	301	GRA
42	US-Los	46.0827	-89.9792	480	WET
43	US-Me2	44.4523	-121.5574	1253	ENF
44	US-Me3	44.3154	-121.6078	1005	ENF
45	US-Me5	44.4372	-121.5668	1188	ENF
46	US-Me6	44.3233	-121.6078	998	ENF
47	US-MMS	39.3232	-86.4131	275	DBF
48	US-MOz	38.7441	-92.2000	219	DBF
49	US-Mpj	34.4385	-106.2377	2196	WSA
50	US-MRf	44.6465	-123.5515	263	ENF
51	US-Myb	38.0498	-121.7651	-1	WET
52	US-NC1	35.8118	-76.7119	5	ENF
53	US-NC2	35.8030	-76.6685	5	ENF
54	US-Ne1	41.1651	-96.4766	361	CRO
55	US-Ne2	41.1649	-96.4701	362	CRO
56	US-Ne3	41.1797	-96.4397	363	CRO
57	US-NR1	40.0329	-105.5464	3050	ENF
58	US-Oho	41.5545	-83.8438	230	DBF
59	US-PFa	45.9459	-90.2723	470	MF
60	US-Ro1	44.7143	-93.0898	290	CRO
61	US-Ro3	44.7217	-93.0893	260	CRO
62	US-Seg	34.3623	-106.7020	1596	GRA
63	US-Ses	34.3349	-106.7442	1604	OSH
64	US-Slt	39.9138	-74.5960	30	DBF
65	US-Snd	38.0373	-121.7537	-5	GRA
66	US-SO2	33.3738	-116.6228	1394	CSH
67	US-SO3	33.3771	-116.6226	1429	CSH
68	US-SO4	33.3845	-116.6406	1429	CSH
69	US-SRC	31.9083	-110.8395	991	OSH
70	US-SRG	31.7894	-110.8277	1291	GRA
71	US-SRM	31.8214	-110.8661	1120	WSA
72	US-Syv	46.2420	-89.3477	540	MF
73	US-Ton	38.4316	-120.9660	177	WSA
74	US-Tw1	38.1074	-121.6469	-9	WET
75	US-Tw3	38.1159	-121.6467	-9	CRO
76	US-Twt	38.1087	-121.6530	-5	CRO
77	US-UMB	45.5598	-84.7138	234	DBF
78	US-UMd	45.5625	-84.6975	239	DBF
79	US-Var	38.4133	-120.9507	129	GRA
80	US-Vcm	35.8884	-106.5321	3030	ENF
81	US-Vcp	35.8642	-106.5967	2500	ENF
82	US-WBW	35.9588	-106.5967	283	DBF
83	US-WCr	45.8059	-90.0799	520	DBF
84	US-Whs	31.7438	-110.0522	1370	OSH
85	US-Wjs	34.4255	-105.8615	1931	SAV
86	US-Wkg	31.7365	-109.9419	1531	GRA
87	US-Wlr	37.5208	-96.8550	408	GRA
88	US-WPT	37.5208	-96.8550	175	WET
89	US-Wrc	45.8205	-121.9519	371	ENF

Curriculum Vitae

Name:	Fang He
Post-Secondary Education and Degrees:	Statistics, Ph.D. University of Western Ontario 2014 - 2018
	Statistics, M.Sc. University of Western Ontario 2013 - 2014
	Mathematics, B.Sc. South China University of Technology 2009 - 2013
Related Work Experience:	Statistical Consultants The University of Western Ontario 2015 - 2018
	Lead R Workshops (at least four times a year) The University of Western Ontario 2015 - 2018